
Downloaded from: http://researchonline.lshtm.ac.uk/2197028/

DOI: https://doi.org/10.17037/PUBS.02197028

Usage Guidelines:

Please refer to usage guidelines at http://researchonline.lshtm.ac.uk/policies.html or alternatively contact researchonline@lshtm.ac.uk.

Available under license: http://creativecommons.org/licenses/by-nc-nd/2.5/
‘Genetic diversity in Trypanosoma cruzi: marker development and applications; natural population structures, and genetic exchange mechanisms’

Louisa Alexandra Messenger

Thesis submitted to the University of London in fulfilment of the requirements for the degree of Doctor of Philosophy.

September 2014

Supervisor: Professor Michael A. Miles

Department of Pathogen Molecular Biology
Faculty of Infectious and Tropical Diseases
London School of Hygiene and Tropical Medicine
United Kingdom

Funded by:
Candidate declaration

I, Louisa Alexandra Messenger, confirm that the work presented in this thesis is my own.

Where information has been derived from other sources, I confirm that this has been indicated.

[Signature]

September, 2014
Abstract

Chagas disease remains the most important parasitic infection in Latin America. The aetiological agent, *Trypanosoma cruzi* (Kinetoplastida: Trypanosomatidae), is a complex vector-borne zoonosis transmitted in the faeces of hematophagous triatomine bugs (Hemiptera: Reduviidae: Triatominae), and maintained by mammalian reservoir hosts ranging from the southern United States to Argentinean Patagonia. In the absence of chemotherapy, infection is life-long and can lead to a spectrum of pathological sequelae ranging from subclinical to lethal cardiac and/or gastrointestinal complications in up to 30% of patients.

*Trypanosoma cruzi* displays remarkable genetic diversity, which has long been suspected to contribute to the considerable variation in clinical symptoms observed between endemic regions. Currently, isolates of *T. cruzi* can be assigned to a minimum of six stable genetic lineages or discrete typing units (DTUs) (TcI-TcVI), which are broadly associated with disparate ecologies, transmission cycles and geographical distributions. The principal mode of reproduction among *T. cruzi* strains is the subject of an intense, decades-old debate. Despite the existence of two recent natural hybrid lineages (TcV and TcVI), which resemble meiotic F1 progeny, a pervasive view is that recombination has been restrained at an evolutionary scale and is of little epidemiological relevance to contemporary parasite populations.

The aim of this PhD project was to investigate *T. cruzi* genetic diversity through significant development of phylogenetic markers and their application to the characterization of natural parasite population structures and genetic exchange mechanisms. Multiple, single-copy, chromosomally-independent, nuclear housekeeping genes were assessed initially for their ability to allocate isolates to DTU-level, to facilitate higher resolution intra-lineage analyses and finally for their inclusion alongside additional targets in a standardized *T. cruzi* multilocus sequence typing (nMLST) scheme. For the immediate future, nuclear MLST, using a panel of four to seven nuclear loci, is a robust, reproducible and highly discriminatory method that has potential to become the new gold standard for *T. cruzi* DTU assignment.

To investigate natural parasite population structures and uncover evidence of genetic exchange, a high resolution mitochondrial MLST (mtMLST) scheme, based on ten gene fragments, was developed and evaluated against current nuclear markers (multilocus microsatellite typing; MLMT) using isolates belonging to the oldest and most widely distributed lineage (TcI). Observations of gross nuclear-mitochondrial phylogenetic incongruence indicate that recombination is ongoing, geographically widespread and continues to influence natural populations, challenging the traditional paradigm of clonality in *T. cruzi*.

Application of this combined nuclear-mitochondrial methodology to intensively sampled, minimally-subdivided TcI populations revealed extensive mitochondrial introgression within a disease focus in North-East Colombia as well as among arboreal transmission cycles in Bolivia. Failure to detect any reciprocal nuclear hybridization among recombinant strains
may be indicative of alternate, cryptic mating strategies in *T. cruzi*, which are challenging to reconcile with both *in vitro* parasexual mechanisms of genetic exchange described, and patterns of Mendelian allele inheritance among natural hybrid DTUs.

High resolution genotyping of Tcl populations was also undertaken to explore the interaction between parasite genetic heterogeneity and ecological biodiversity, exposing the significant impact human activity has had on *T. cruzi* evolution. Reduced genetic diversity, accelerated parasite dissemination between densely populated areas and mitochondrial gene flow between domestic and sylvatic populations, suggests humans may have played a crucial role in *T. cruzi* dispersal across the Bolivian highlands. Parallel reductions in genetic diversity were observed among isolates from the Brazilian Atlantic Forest, attributable to ongoing anthropogenic habitat fragmentation. By comparison domestic Tcl isolates (TcIDOM) are divergent from their sylvatic counterparts, but also genetically homogeneous, and likely to have originated in North/Central America before distribution southwards. Molecular dating of Colombian TclDOM clones confirmed that this clade emerged 23,000 ± 12,000 years, coinciding with the earliest human migration into South America.

Lastly, Illumina amplicon deep sequencing markers were developed to explore the interaction between parasite multiclontality and clinical status of chronic Chagas disease. An unprecedented level of intra-host genetic diversity was detected, highlighting putative diversifying selection affecting antigenic surface proteases, which may facilitate survival in the mammalian host. In lieu of comparative genomics of representative *T. cruzi* field isolates, not yet a reality, as is the case with other more experimentally-tractable trypanosomatids, presented herein are some of the highest resolution genotyping techniques developed in *T. cruzi* to date, which have the potential to expand our current understanding of parasite genetic diversity and its relevance to clinical outcome of Chagas disease.
### Table of Contents

Candidate declaration ............................................................................................................. 2
Abstract ................................................................................................................................. 3
Acknowledgements .................................................................................................................. 8
List of figures ........................................................................................................................... 9
List of tables ............................................................................................................................ 11
Abbreviations .......................................................................................................................... 12
Candidate’s contributions to publications ............................................................................. 16
1. Introduction ....................................................................................................................... 20
   1.1 Trypanosoma cruzi life cycle and transmission ............................................................ 20
   1.2 Epidemiology and public health importance of Chagas disease ................................. 22
   1.3 Diagnosis and clinical manifestations of vector-borne Chagas disease ..................... 23
   1.4 Diagnosis and clinical manifestations of congenital Chagas disease ......................... 26
   1.5 Pathogenesis and immunology of Chagas disease ....................................................... 27
   1.6 Current treatment and management of Chagas disease ............................................. 30
   1.7 Prospective treatment of Chagas disease ................................................................. 30
   1.8 Vector species, behaviours and ecologies ................................................................. 31
   1.9 Mammalian reservoir hosts and transmission cycle dynamics ................................. 34
   1.10 Control of Chagas disease ....................................................................................... 36
   1.11 The origins of T. cruzi ............................................................................................ 38
   1.12 Contemporary population structure of T. cruzi ....................................................... 40
   1.13 Molecular epidemiology of T. cruzi .......................................................................... 41
   1.14 Genetic exchange in kinetoplastid protozoa and its phenotypic implications ......... 44
   1.15 PhD aims, biological objectives and milestones ....................................................... 47
2. Materials and methods ....................................................................................................... 48
   2.1 Ethics statements ....................................................................................................... 48
   2.3.1 Parasite (epimastigote) culture .............................................................................. 48
   2.3.2 Isolating parasite field strains ............................................................................... 49
   2.3.3 Biological cloning of parasites on solid media ...................................................... 49
   2.3.4 Extraction of parasite genomic DNA .................................................................... 50
   2.3.5 Polymerase chain reaction (PCR) ......................................................................... 50
   2.3.6 Agarose gel electrophoresis .................................................................................. 50
   2.3.7 PCR purification ..................................................................................................... 50
   2.3.8 Chain termination DNA sequencing .................................................................... 51
   2.3.9 Ligation reactions .................................................................................................. 51
   2.3.10 Preparation of chemically competent bacterial cells .......................................... 51
Acknowledgements

First and foremost I would like to thank my supervisor Prof. Michael Miles for his encouragement, inspiration and constant support. Without the guidance, training, and enthusiasm from Dr. Martin Llewellyn, Dr. Michael Lewis, Dr. Matthew Yeo and Tapan Bhattacharyya this project would not have been possible.

I have been incredibly fortunate to receive parasite DNA from members of the European Union Seventh Framework Programme ChagasEpiNet: Dr. Lineth García and Dr. Faustino Torrico (Universidad Mayor de San Simón, Bolivia), Dr. Juan-David Ramírez and Prof. Felipe Guhl, (Universidad de los Andes, Colombia), Maikell Segovia and Prof. Hernan Carrasco (Universidad Central de Venezuela, Venezuela), Dr. Patricio Diosque (Universidad Nacional de Salta, Argentina), Valdirene dos Santos Lima and Prof. Ana Jansen (Fundação Oswaldo Cruz Foundation, Brazil) and Federico Zumaya-Estrada (Instituto Nacional de Salud Pública, México).

Mitochondrial genome data and training in bioinformatics were kindly provided by Prof. Björn Andersson and Dr. Oscar Franzén at the Karolinska Institutet, Sweden.

My many thanks to Dr. Caryn Bern, Prof. Robert Gilman and Dr. Lineth García for hosting me during my adventures in Bolivia.

This PhD was funded by a Doctoral Training Grant from the Biotechnology and Biological Sciences Research Council (BBSRC). Funding for fieldwork to Bolivia was generously awarded from the Chadwick Trust Travelling Scholarship (University College London), the Dr. Gordon Smith Travelling Scholarship (LSHTM), and the Royal Society of Tropical Medicine and Hygiene, London.
List of figures

**Figure 1.1.** *Trypanosoma cruzi* morphological forms.

**Figure 1.2.** The life cycle of *T. cruzi*.

**Figure 1.3.** The distribution of endemic and non-endemic Chagas disease.

**Figure 1.4.** Schematic of the clinical progression of Chagas disease.

**Figure 1.5.** Acute and chronic clinical symptoms of Chagas disease.

**Figure 1.6.** Venous blood draw from a one month year old participant enrolled in an ongoing congenital Chagas disease surveillance study in Hospital Municipal Camiri, Santa Cruz department, Bolivia.

**Figure 1.7.** Immune responses in acute Chagas disease.

**Figure 1.8.** Immunoregulatory model in chronic indeterminate and cardiac Chagas disease.

**Figure 1.9.** Major triatomine bug species responsible for *T. cruzi* transmission.

**Figure 1.10.** Geographical distribution of the 14 most important vector species of Chagas disease.

**Figure 1.11.** Examples of important sylvatic reservoirs of *T. cruzi* infection: *Didelphis albiventris* and *Dasypus novemcinctus*.

**Figure 1.12.** Example of domestic transmission in an endemic region of the Gran Chaco, Santa Cruz department, Bolivia.

**Figure 1.13.** Example of potential peridomestic reservoir hosts in an endemic region of the Gran Chaco, Santa Cruz department, Bolivia.

**Figure 1.14.** The seroprevalence of Chagas disease and geographical distribution of control programmes.

**Figure 1.15.** Two hypotheses for the evolution of the *T. cruzi* clade: (a) the southern super-continent hypothesis and (b) the bat seeding hypothesis.

**Figure 1.16.** Comparison of the two-hybridization and three ancestor models which describe the role of genetic exchange during the evolution of *T. cruzi*.

**Figure 1.17.** Approximate geographical distributions of *T. cruzi* lineages and transmission cycles.
Figure 1.18. Mechanisms of genetic exchange among selected parasitic protozoa, predicted implications for population structure and observed population structures in the field.

Figure 2.1. Recommended triple-assay for discriminating *T. cruzi* DTUs.

Figure 2.2. Examples of PCR-RFLP genotyping profiles.

Figure 2.3. Example of a heterozygous sequence chromatogram.

Figure 2.4. Microsatellite primer positions in a 96-well plate.

Figure 2.5. Schematic of multiplexing microsatellite PCR products.

Figure 2.6. Example of an ambiguous microsatellite profile for strains from different *T. cruzi* DTUs.
List of tables

Table 1.1. An overview of ecotopes, sylvatic vectors.getHosts, geographical distributions and clinical associations of the major *T. cruzi* DTUs.

Table 2.1. PCR-RFLP gene fragments and primer details.

Table 2.2. *T. cruzi* genotype assignment of PCR amplification product sizes (bp).

Table 2.3. *T. cruzi* nuclear gene fragments and primer details.

Table 2.4. *T. cruzi* maxicircle gene fragments and primer details.

Table 2.5. *T. cruzi* microsatellite loci and primer details.

Table 3.1 Comparison of *T. cruzi* historical and contemporary nomenclatures.

Table 3.2. Overview of current and historical *T. cruzi* genotyping methods.

Table 5.1. Summary of field evidence of genetic exchange in *T. cruzi*. 


<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>aCSDI</td>
<td>Absolute chromosomal size difference index</td>
</tr>
<tr>
<td>AIC</td>
<td>Akaike information criterion</td>
</tr>
<tr>
<td>AMCHA</td>
<td>Initiative of the Amazon Countries for Surveillance and Control of Chagas Disease</td>
</tr>
<tr>
<td>AMOVA</td>
<td>Analysis of molecular variance</td>
</tr>
<tr>
<td>A_r</td>
<td>Allelic richness</td>
</tr>
<tr>
<td>BBSRC</td>
<td>Biotechnology and Biological Sciences Research Council</td>
</tr>
<tr>
<td>BIC</td>
<td>Bayesian information criterion</td>
</tr>
<tr>
<td>BEAST</td>
<td>Bayesian evolutionary analysis by sampling trees</td>
</tr>
<tr>
<td>BENEFIT</td>
<td>Benznidazole Evaluation for Interrupting Trypanosomiasis</td>
</tr>
<tr>
<td>bp</td>
<td>Base-pair(s)</td>
</tr>
<tr>
<td>BSA</td>
<td>Bovine serum albumin</td>
</tr>
<tr>
<td>BSP</td>
<td>Bayesian skyline plot</td>
</tr>
<tr>
<td>DALY</td>
<td>Disability-adjusted life years</td>
</tr>
<tr>
<td>DAPC</td>
<td>Discriminant analysis of principal components</td>
</tr>
<tr>
<td>D_A_S</td>
<td>Distance allele shared</td>
</tr>
<tr>
<td>ddH_2O</td>
<td>Double-distilled water</td>
</tr>
<tr>
<td>DHFR-TS</td>
<td>Dihydrofolate reductase-thymidylate synthase</td>
</tr>
<tr>
<td>DNA</td>
<td>Deoxyribonucleic acid</td>
</tr>
<tr>
<td>dNTP</td>
<td>Deoxyribonucleoside triphosphate</td>
</tr>
<tr>
<td>DST</td>
<td>Diploid sequence type</td>
</tr>
<tr>
<td>DTU</td>
<td>Discrete typing unit</td>
</tr>
<tr>
<td>ECG</td>
<td>Electrocardiogram</td>
</tr>
<tr>
<td>EDTA</td>
<td>Ethylenediaminetetraacetic acid disodium salt C_{10}H_{14}N_{2}Na_{2}O_{8}</td>
</tr>
<tr>
<td>ELISA</td>
<td>Enzyme-linked immunosorbent assay</td>
</tr>
<tr>
<td>EtBr</td>
<td>Ethidium bromide C_{21}H_{20}BrN_{3}</td>
</tr>
<tr>
<td>Abbreviation</td>
<td>Description</td>
</tr>
<tr>
<td>--------------</td>
<td>-------------</td>
</tr>
<tr>
<td>FACS</td>
<td>Fluorescence activated cell sorting</td>
</tr>
<tr>
<td>FBS</td>
<td>Fetal bovine serum</td>
</tr>
<tr>
<td>FFLB</td>
<td>Fluorescent fragment length barcoding</td>
</tr>
<tr>
<td>GEB</td>
<td>Guanidine hydrochloride-EDTA</td>
</tr>
<tr>
<td>HRM</td>
<td>High resolution melting</td>
</tr>
<tr>
<td>IAM</td>
<td>Infinite alleles model</td>
</tr>
<tr>
<td>IBD</td>
<td>Isolation by distance</td>
</tr>
<tr>
<td>ICD</td>
<td>Implantable cardioverter defibrillator</td>
</tr>
<tr>
<td>IFA</td>
<td>Indirect immunofluorescence assay</td>
</tr>
<tr>
<td>IFN-γ</td>
<td>Interferon-gamma</td>
</tr>
<tr>
<td>IHA</td>
<td>Indirect hemagglutination assay</td>
</tr>
<tr>
<td>IL</td>
<td>Interleukin</td>
</tr>
<tr>
<td>INCOSUR</td>
<td><em>Iniciativa del Cono Sur para Control/Eliminación de Chagas</em> (Southern Cone Initiative for Control/Elimination of Chagas Disease)</td>
</tr>
<tr>
<td>IPA</td>
<td><em>Iniciativa de los Países Andinos</em> (Initiative of the Andean Countries)</td>
</tr>
<tr>
<td>IPCA</td>
<td><em>Iniciativa de los Países de Centro América para la Interrupción de la Transmisión Vectorial y Transfusional de la Enfermedad de Chagas</em> (Initiative of the Central American Countries for the Interruption of the Vectorial and Transfusional Transmission of Chagas Disease)</td>
</tr>
<tr>
<td>IPTG</td>
<td>Isopropyl β-D-1-thiogalactopyranoside</td>
</tr>
<tr>
<td>IRB</td>
<td>Institutional review board</td>
</tr>
<tr>
<td>IUPAC</td>
<td>International Union of Pure and Applied Chemistry</td>
</tr>
<tr>
<td>kDNA</td>
<td>Kinetoplast DNA</td>
</tr>
<tr>
<td>KH</td>
<td>Kishino-Hasegawa</td>
</tr>
<tr>
<td>LB</td>
<td>Lysogeny broth</td>
</tr>
<tr>
<td>LD</td>
<td>Linkage disequilibrium</td>
</tr>
<tr>
<td>LIT</td>
<td>Liver infusion tryptose</td>
</tr>
<tr>
<td>Abbreviation</td>
<td>Description</td>
</tr>
<tr>
<td>--------------</td>
<td>-------------</td>
</tr>
<tr>
<td>LMP</td>
<td>Low melting point</td>
</tr>
<tr>
<td>LOH</td>
<td>Loss of heterozygosity</td>
</tr>
<tr>
<td>LSHTM</td>
<td>London School of Hygiene and Tropical Medicine, UK</td>
</tr>
<tr>
<td>LSSP</td>
<td>Low-stringency single specific primer</td>
</tr>
<tr>
<td>MCMC</td>
<td>Markov chain-Monte Carlo</td>
</tr>
<tr>
<td>Met-II</td>
<td>Metacyclin-II</td>
</tr>
<tr>
<td>Met-III</td>
<td>Metacyclin-III</td>
</tr>
<tr>
<td>mHVR</td>
<td>Minicircle hypervariable region</td>
</tr>
<tr>
<td>ML</td>
<td>Maximum-likelihood</td>
</tr>
<tr>
<td>MLD</td>
<td>Multilocus linkage disequilibrium</td>
</tr>
<tr>
<td>MLG</td>
<td>Multilocus genotype</td>
</tr>
<tr>
<td>MLEE</td>
<td>Multilocus enzyme electrophoresis</td>
</tr>
<tr>
<td>MLMT</td>
<td>Multilocus microsatellite typing</td>
</tr>
<tr>
<td>MLST</td>
<td>Multilocus sequence typing</td>
</tr>
<tr>
<td>MRCA</td>
<td>Most recent common ancestor</td>
</tr>
<tr>
<td>mtMLST</td>
<td>Maxicircle multilocus sequence typing</td>
</tr>
<tr>
<td>MYA</td>
<td>Million years ago</td>
</tr>
<tr>
<td>NJ</td>
<td>Neighbour-joining</td>
</tr>
<tr>
<td>nMLST</td>
<td>Nuclear multilocus sequence typing</td>
</tr>
<tr>
<td>NO</td>
<td>Nitric oxide</td>
</tr>
<tr>
<td>PA</td>
<td>Private allele</td>
</tr>
<tr>
<td>PAHO</td>
<td>Pan American Health Organization</td>
</tr>
<tr>
<td>PAMP</td>
<td>Pathogen-associated molecular pattern</td>
</tr>
<tr>
<td>PCE</td>
<td>Preponderate clonal evolution</td>
</tr>
<tr>
<td>PCR</td>
<td>Polymerase chain reaction</td>
</tr>
<tr>
<td>PFGE</td>
<td>Pulsed-field gel electrophoresis</td>
</tr>
<tr>
<td>qPCR</td>
<td>Quantitative PCR</td>
</tr>
<tr>
<td>Abbreviation</td>
<td>Definition</td>
</tr>
<tr>
<td>--------------</td>
<td>------------</td>
</tr>
<tr>
<td>RAPD</td>
<td>Random amplification of polymorphic DNA</td>
</tr>
<tr>
<td>RB19</td>
<td>RNA-binding-protein-19</td>
</tr>
<tr>
<td>RFLP</td>
<td>Restriction fragment length polymorphism</td>
</tr>
<tr>
<td>SD</td>
<td>Standard deviation</td>
</tr>
<tr>
<td>SH</td>
<td>Shimodaira-Hasegawa</td>
</tr>
<tr>
<td>SL-IR</td>
<td>Spliced leader intergenic region</td>
</tr>
<tr>
<td>SNP</td>
<td>Single nucleotide polymorphism</td>
</tr>
<tr>
<td>SOC</td>
<td>Super optimal broth with catabolite repression</td>
</tr>
<tr>
<td>ST</td>
<td>Sequence type</td>
</tr>
<tr>
<td>TAE</td>
<td>Tris-acetic acid-EDTA buffer</td>
</tr>
<tr>
<td>TcAPX</td>
<td>Ascorbate-dependent haemoperoxidase</td>
</tr>
<tr>
<td>TcGPXII</td>
<td>Glutathione-dependent peroxidase II</td>
</tr>
<tr>
<td>TcMPX</td>
<td>Mitochondrial peroxidase</td>
</tr>
<tr>
<td>TNF-α</td>
<td>Tumor necrosis factor-alpha</td>
</tr>
<tr>
<td>TR</td>
<td>Trypanothione reductase</td>
</tr>
<tr>
<td>TRAENA</td>
<td>Tratamiento con Benznidazol en pacientes Adultos con Enfermedad de Chagas Crónica</td>
</tr>
<tr>
<td>UV</td>
<td>Ultraviolet</td>
</tr>
<tr>
<td>VSG</td>
<td>Variant surface glycoprotein</td>
</tr>
<tr>
<td>WGA</td>
<td>Whole genome amplification</td>
</tr>
<tr>
<td>WGS</td>
<td>Whole genome sequencing</td>
</tr>
<tr>
<td>WHO</td>
<td>World Health Organization</td>
</tr>
<tr>
<td>X-gal</td>
<td>5-bromo-4-chloro-3-indolyl-β-D-galactopyranoside</td>
</tr>
</tbody>
</table>
Candidate’s contributions to publications

Chapter 3


The candidate was involved in study design, marker choice and was responsible for complete characterization of four of the nine housekeeping genes under evaluation (LYT1, Met-II, Met-III and RB19). The candidate also participated in data interpretation and assisted in manuscript drafting.


The candidate generated complete sequence data for five of thirteen housekeeping genes under evaluation (Met-II, Met-III, RB19, TcAPX and TcMPX) and participated in data interpretation, analysis and initial manuscript drafting.


The candidate designed the study and performed all mitochondrial sequencing including assembly of the Sylvio X10/1 mitochondrial genome with guidance from collaborators at the Karolinska Institutet, Sweden. The candidate also generated the microsatellite data, analyzed all data, drafted the manuscript and was responsible for final manuscript revisions for publication.

The candidate participated in mitochondrial sequencing marker design, provided unpublished reference datasets, laboratory samples, coordinated provision of Bolivian field samples and contributed to data analysis, interpretation and manuscript drafting.

Chapter 4


The candidate designed the study in close collaboration with Dr. Martin Llewellyn, derived the majority of biological clones analyzed, performed all of the genotyping experiments, analyzed the data and drafted the manuscript.


The candidate was responsible for generating, assembling and analyzing all mitochondrial sequence data. The candidate also participated in overall data analysis and drafting of the final manuscript.


The candidate had significant input into experimental design, contributed laboratory reagents, supervision and reference datasets. The candidate also participated in mitochondrial data generation, analysis and drafting of the final manuscript.

Chapter 5


The candidate had significant input into experimental design, contributed laboratory reagents and reference datasets, supervised the generation of mitochondrial data, and participated in data analysis and drafting of the final manuscript.

The candidate designed the study, performed all of the genotyping experiments, analyzed the data and drafted the manuscript.

**Appendix 1**


The candidate drafted the manuscript in close consultation with the other co-authors.


The candidate was responsible for assembling the whole maxicircle genome sequence for *T. c. marinkellei* and performing all associated mitochondrial analyses. The candidate drafted all sections pertaining to mitochondrial data for the final manuscript.


The candidate participated in data assembly, analysis, interpretation and manuscript preparation.


The candidate was responsible for generating, assembling and analyzing all mitochondrial sequence data. The candidate also participated in overall data interpretation and drafting of the final manuscript.

The candidate contributed laboratory reagents (live epimastigote and metacyclic parasites) used to generate parasite lysates and infected mouse sera, and technical support during data generation.
1. Introduction

1.1 *Trypanosoma cruzi* life cycle and transmission

*Trypanosoma cruzi*, the aetiological agent of Chagas disease (American trypanosomiasis) is a hemoflagellate protozoan parasite belonging to the genus *Trypanosoma* within the class Kinetoplastida. This group also includes other human pathogens of public health importance, notably *Trypanosoma brucei*, the causative agent of sleeping sickness (African trypanosomiasis), and *Leishmania* species. Kinetoplastids are characterized by a number of unique biological features including, extra-nuclear DNA (kDNA) in the form of minicircles and maxicircles (functionally equivalent to eukaryotic mitochondria), RNA editing and almost exclusive post-transcriptional gene regulation (Simpson *et al.*, 1988; Lukes *et al.*, 2002; Clayton and Shapira, 2007).

*T. cruzi* is a complex vector-borne zoonosis transmitted by hematophagous triatomine bugs (Hemiptera: Reduviidae: Triatominae) and maintained by a wide range of mammalian reservoirs. In order to successfully colonize both the mammalian host and insect vector, *T. cruzi* assumes three distinct morphological forms at different developmental stages of its life cycle (Figure 1.1) (Tyler and Engman, 2001; 2003).

![Figure 1.1. *Trypanosoma cruzi* morphological forms](http://www.stanford.edu, http://winona.edu, http://www.dpd.cdc.gov)

Amastigote and epimastigote forms (Figure 1.1 A and C) replicate by binary fission within phagocytic and non-phagocytic mammalian cells and in the hindgut of the triatomine bug vector, respectively. Trypomastigote forms (Figure 1.1 B) are non-replicative and are present at two distinct life cycle stages: (i) in the bloodstream of the mammalian host (bloodstream form trypomastigotes) and (ii) in the rectum and faeces of triatomine bug vectors (infective metacyclic trypomastigotes).

The life cycle of *T. cruzi* is illustrated in Figure 1.2. Infective metacyclic trypomastigotes are deposited on the skin of the mammalian host in faecal droplets extruded by a blood-feeding triatomine bug. Parasites enter the host via contamination of the bite wound or alternatively through skin abrasions, exposed mucous membranes or the conjunctiva. Once inside the host, motile trypomastigotes are able to invade a range of phagocytic and non-phagocytic nucleated cells via both lysosomal-dependent and independent mechanisms (reviewed by Epting *et al.*, 2010; Caradonna and Burleigh, 2011).
Upon cell entry, the parasite is taken up into a membrane-bound (parasitophorous) vacuole which subsequently fuses with a lysosome; exposure to decreasing pH stimulates parasite differentiation to the intracellular amastigote form and its concomitant release into the cytosol. Here, amastigotes multiply asexually to form pseudocysts, which can arise in a variety of host tissues but predominate in cardiac, smooth and skeletal muscles and reticuloendothelial cells in the liver, spleen and lymphatic system. Within pseudocysts, amastigotes differentiate into non-replicative, flagellated trypomastigotes that, upon cell lysis, can either infect adjacent tissues or disseminate throughout the bloodstream or lymph.

Triatomine bugs feeding on an infected mammal may ingest extracellular trypomastigotes, which pass to the midgut where transformation to the amastigote form occurs. Differentiation of amastigotes into epimastigotes occurs in response to decreasing environmental glucose levels from the mammalian host to the insect vector (Tyler and Engman, 2000). Epimastigotes multiply by binary fission in the hindgut and migrate to the rectum where they attach hydrophobically to the waxy gut cuticle by their flagella and transform into infective metacyclic trypomastigotes, thus completing the life cycle.

**Figure 1.2.** The life cycle of *T. cruzi*. Red and blue arrows indicate parasite developmental stages in the insect vector and human host, respectively. Source: [http://www.dpd.cdc.gov/dpdx](http://www.dpd.cdc.gov/dpdx).
1.2 Epidemiology and public health importance of Chagas disease

Chagas disease remains the most important parasitic infection in Latin America. It is estimated that 8-10 million individuals are infected, with a further 90 million at risk (Hotez et al., 2008; Rassi Jr et al., 2010). T. cruzi has a broad geographical range that extends from the southern United States to Argentinean Patagonia (latitude 42°N to latitude 40°S; Figure 1.3). However, human disease transmission is principally restricted to areas where individuals are exposed to the infected faeces of domiciliated or invasive triatomine bugs. In endemic countries, Chagas disease disproportionally affects impoverished communities resulting in the loss of 0.7 million Disability Adjusted Life Years (DALYs) (WHO, 2008) and 12,500 deaths per annum (Lescure et al., 2010).

Figure 1.3. The distribution of endemic (light grey) and non-endemic Chagas disease (see colour key) estimated by Rassi Jr et al., 2010. Source: Rassi Jr et al., 2010.

While the majority of T. cruzi cases are vector-borne, oral infection is an important secondary transmission route which is often responsible for regional microepidemics of acute Chagas disease in areas devoid of domestic triatomine species, such as the Amazon Basin (recently reviewed by Shikanai-Yasuda and Carvalho, 2012). Ingestion of food/drink, contaminated with triatomine faeces, is generally associated with massive parasitic infection resulting in more severe acute clinical presentations and higher mortality rates (Yoshida, 2009; Alarcón de Noya et al., 2010).

In recent years, a significant proportion of the infected population has emigrated from rural areas, leading to the urbanization of Chagas disease in endemic countries as well as internationally (Gürtler, 2009). Chagas disease is now considered an emergent global public health problem associated with congenital transmission (Bern et al., 2011a), blood transfusions (Jackson et al., 2010a) and organ transplantations (Centers of Disease Control, 2006; Kransdorf et al., 2013) (reviewed by Schmunis and Yadon, 2010; Gascon et al., 2010). Current estimates indicate 1-26% of migrants are T. cruzi seropositive, depending on the
country and/or the migrants’ nationality (Jackson et al., 2009). In the United States over 300,000 individuals are estimated to harbor *T. cruzi*, forming a potential parasite reservoir for autochthonous transmission (Bern and Montgomery, 2009; Bern et al., 2011b) (Figure 1.3).

Lastly, *T. cruzi* infection can also result from occupational exposure, with at least sixty-five cases of laboratory-acquired Chagas disease reported to date (Herwaldt, 2001).

1.3 Diagnosis and clinical manifestations of vector-borne Chagas disease

In the absence of chemotherapy, infection with *T. cruzi* is life-long and can lead to a spectrum of pathological sequelae ranging from subclinical to debilitation and death (Prata, 2001). Disease progression and clinical outcome are determined by a number of factors including host (recently reviewed by Ayo et al., 2013) and parasite genetics (Campbell et al., 2004), mixed/super-infections (Bustamente et al., 2002; Torrico et al., 2006; Schofield et al., 2006) and geographical strain variation (Miles et al., 1981a; reviewed by Macedo et al., 2004).

Chagas disease can typically be categorized into three stages (Figure 1.4). An initial acute phase begins 4-15 days after infection and lasts up to 8 weeks, during which bloodstream form trypomastigotes are detectable in the peripheral blood. Most individuals remain asymptomatic, while 15-30% present overt clinical manifestations including prolonged fever, malaise, lymphadenopathy and hepatosplenomegaly. Inflammation and parasite multiplication at the site of inoculation can develop into an oedematous nodule or chagoma (Figure 1.5). In addition, up to 50% of cases present with conjunctivitis and unilateral palperbral oedema (Romaña’s sign). Mortality during the acute stage is low (<5-10% of cases) and predominates in children under five and immunocompromised individuals, usually resulting from severe myocarditis, meningoencephalitis, or both (Rassi Jr et al., 2010). The acute phase resolves spontaneously in approximately 90% of patients, even if the infection is left untreated.
Following the acute phase, 60-70% of infected individuals progress to an asymptomatic or indeterminate period of clinical latency, which can persist indefinitely. Indeterminate Chagas disease is characterized by *T. cruzi* seropositivity, a normal 12-lead electrocardiogram (ECG) and normal radiological examination of the chest, oesophagus and colon (Rassi Jr *et al.*, 2010). During this stage parasitaemia is undetectable by direct microscopy and diagnosis is reliant on high levels of circulating anti-*T. cruzi* IgG detected by at least two different serological methods (usually enzyme-linked immunosorbent assay; ELISA, indirect immunofluorescence; IFA or indirect haemagglutination; IHA) (Duarte *et al.*, 2014).

The remaining 30-40% of patients develop chronic Chagas disease within 10-30 years post-infection at a rate of approximately 2% per year (Prata, 2001). Direct progression from the acute phase to the chronic form of Chagas disease has been reported in a minority of individuals. In addition, reactivation of Chagas disease can occur in chronically infected patients who become immunocompromised by immunosuppressive treatment or co-infection with HIV (Bern, 2012).
Chronic Chagas disease is characterised by irreversible lesions to the cardiac or digestive tissues or, in rare cases, to both. Early chagasic cardiomyopathy typically presents as conduction system abnormalities, particularly right bundle branch block, left anterior fascicular block and premature left ventricular wall contractions (Maguire et al., 1987). More advanced manifestations include ventricular tachycardia, high degree atrioventricular block and progressive dilated cardiomyopathy with congestive heart failure (Rassi Jr et al., 2000). Sudden death accounts for nearly two-thirds of chagasic cardiomyopathy-related mortality and can affect both patients with end-stage heart disease, as well as those who were previously asymptomatic (Rassi Jr et al., 2001). Denervation, muscular hypertrophy and luminal dilatation of the upper and lower alimentary tract can lead to the development of megaesophagus or megacolon, respectively (De Oliveira et al., 1998) (Figure 1.5).

**Figure 1.5.** Acute and chronic clinical symptoms of Chagas disease: A – Romaña’s sign; B – Chagoma; C – Apical aneurysm; D – Chagasic megaesophagus; E – Chagasic megacolon.
Sources: WHO, 2002; Kinoshita-Yanaga et al., 2009; Coura and Vinas, 2010 and Rassi Jr et al., 2010.

In general, chagasic megasyndromes are rarer than cardiac but the prevalence of different clinical forms of Chagas disease varies considerably between individuals and endemic regions (Miles et al., 1981a; Rezende, 1976). Chronic Chagas disease pathology demonstrates geographically-restricted clinical profiles; cardiomyopathies occur throughout Central and South America, whereas gastrointestinal complications are reported almost
exclusively from south of the Amazon Basin. It has been suggested that this geographical heterogeneity is associated with genetic variation among parasite strains (section 1.13) (Miles et al., 1981a; reviewed by Campbell et al., 2004; Macedo et al., 2004). However, the relationship between parasite genotype and clinical outcome remains elusive.

1.4 Diagnosis and clinical manifestations of congenital Chagas disease

With improved vector control, congenital transmission is becoming proportionately more important among chronically-infected populations and now accounts for 25% of new Chagas disease cases (Jannin and Salvatella, 2006). In certain areas of Bolivia, which has the highest T. cruzi seroprevalence (~6.75% of the population), between 72.7-97.1% of adults are infected, including 20-50% of women of child-bearing age (Samuels et al., 2013). Women infected as children remain at risk of vertical transmission for at least 30 years, even if vector-borne transmission were interrupted today (Schmunis and Cruz, 2005). Congenitally infected women can also transmit to their children, sustaining the cycle across generations in the absence of the vector (Carlier and Torrico, 2003).

Diagnosis of congenital Chagas disease involves microscopic observation of trypomastigotes in cord or peripheral blood by microhaematocrit during the first month of life (Freilij et al., 1983; La Fuente et al., 1984) and/or serological detection of anti-T. cruzi IgG at 6-9 months, once maternal antibodies have cleared (Freilij and Altcheh, 1995; Gomes et al., 2009). PCR-based detection of T. cruzi DNA is highly sensitive and can facilitate earlier diagnosis (Diez et al., 2008), but remains only a research tool in endemic areas, due to the need for specialized laboratory capacity, high cost reagents and problems of standardization between laboratories and countries. Although many hospitals have now initiated congenital screening, current diagnostic methods, even when optimally executed, fail to detect over half of infected neonates (Bern et al., 2009) and <20% of at risk infants complete the requisite 9 month follow-ups in programme evaluations (Blanco et al., 2000; Carlier and Torrico, 2003; Alonzo-Vega et al., 2013).

Congenital T. cruzi infection causes a spectrum of clinical manifestations. Disease severity may be related in part to the period of pregnancy during which transmission occurs; transmission early in pregnancy appears to increase risk of spontaneous abortion (Bittencourt and Barbosa 1972), while infection after 22 weeks of gestation is more likely to result in a late stillbirth or an infected live-born infant (Bittencourt et al., 1974; Azogue et al., 1985). For the latter, clinical symptoms range from low birth weight, prematurity and low Apgar scores to meningoencephalitis, hepatosplenomegaly, anaemia, thrombocytopenia and respiratory distress syndrome; mortality occurs in <5-20% of live-born symptomatic infants (Bittencourt et al., 1981; Torrico et al., 2004). Congenital infection also carries a 20-30% risk of chronic cardiac and/or gastrointestinal disease decades later.

Treatment during infancy is significantly more effective and better tolerated than later (Schijman, 2006), but paediatric diagnosis and allocation of finite disease management resources are further complicated by highly variable congenital transmission rates, ranging between 1% to >15% among study populations (Mora et al., 2005; Torrico et al., 2004; 2006;
Factors associated with higher risk of congenital transmission include younger maternal age (presumed to reflect more recent maternal infection) (Torrico et al., 2004), maternal and neonatal immunological responses (Vekemans et al., 2000; Hermann et al., 2004), higher maternal parasitaemia (Bern et al., 2009), HIV and other immunodeficiencies (Freilij and Altcheh, 1995) and in some animal models, parasite strain (Andrade, 1982); evidence for the latter in humans is more equivocal (Virreira et al., 2006; 2007; Corrales et al., 2009).

Figure 1.6. Venous blood draw from a one month year old participant enrolled in an ongoing congenital Chagas disease surveillance study in Hospital Municipal Camiri, Santa Cruz department, Bolivia. Source: L.A. Messenger (LSHTM).

1.5 Pathogenesis and immunology of Chagas disease
Historically, the pathogenesis of Chagas disease, particularly the relative involvement of the parasite and the host immune system to pathology during the chronic stage, have been widely debated (reviewed by Machado et al. 2012). Failure to observe parasites in chronic tissue lesions and the presence of auto-antibodies (Cunha-Neto et al., 1995), potentially resulting from bystander activation and/or molecular mimicry to parasite antigens (Leon and Engman, 2003), led to the ‘autoimmunity hypothesis’ of Chagas disease (Kierszenbaum, 1999) and called into question the benefit of anti-parasitic treatment for disease management (Kierszenbaum, 2005).

This theory has been subsequently challenged by many, based on failure to reproduce chagasic pathology by passive autoantibody transfer and detection of T. cruzi in tissue lesions by PCR (Kierszenbaum, 2003; 2005); although recent mouse data still suggest that pathology
can proceed in the absence of detectable local parasite load (Lewis et al., 2014). Current knowledge indicates that parasite persistence, coupled with an unbalanced immune response in some individuals (which can include autoimmune reactions), is responsible for prolonged inflammatory responses in chronically-infected tissues and resulting pathology, implying elimination of *T. cruzi* may be prerequisite to arrest irreversible disease progression (Dutra and Gollob, 2008).

During the acute phase, tissue damage directly results from the parasite and the host’s robust immunoinflammatory response to its presence (reviewed by Talvani and Teixeira, 2011; Andrade et al., 2014). Macrophages and dendritic cells are activated by pathogen-associated molecular patterns (PAMPs) to secrete interleukin-12 (Campos and Gazzinelli, 2004). IL-12 induces interferon (IFN)-γ synthesis, which augments the production of IL-12 itself and tumor necrosis factor (TNF)-α and polarizes CD4⁺ and CD8⁺ T-cells to produce interferon (IFN)-γ (Martin and Tarleton, 2004). Macrophages activated by TNF-α and IFN-γ play a protective role through the synthesis of trypanocidal nitric oxide (NO) (Chandra et al., 2002; Silva et al., 2003). IFN-γ also stimulates the recruitment of T-cells by inducing expression of pro-inflammatory chemokines and adhesion molecules (reviewed by Teixeira et al., 2002). The acute Th1-predominant immune response is modulated by the production of IL-10 and IL-17 (Hölscher et al., 2000; da Matta Guedes et al., 2010). Acute symptoms resolve spontaneously as parasitaemia decreases to subpatent levels, but without treatment, total parasite clearance virtually never occurs (Rassi Jr et al., 2010).

![Figure 1.7. Immune responses in acute Chagas disease. Source: Andrade et al., 2014.](image)
During chronic infection, adequate T-cell-mediated immunity is essential to control parasitaemia (Reis et al., 1993; Dutra et al., 1994; da Silveira et al., 2007); immunosuppression can lead to lethal reactivation (Cordova et al., 2008; Bacal et al., 2010; Bern, 2012). However, concomitant failure to downregulate the inflammatory response, triggered by parasite persistence, plays a key role in cardiac pathogenesis (Dutra et al., 2005; Martin-Neto et al., 2007; Dutra and Golllob, 2008) (Figure 1.8). Comparisons between asymptomatic and cardiac patients, indicate the former display higher expressions of IL-10 (Souza et al., 2004), CD25^{High}CD4^{+} regulatory T-cells (de Araújo et al., 2011) and surface expression of CTLA-4 by CD8^{+} T-cells (Souza et al., 2007), which may all contribute to control of cytolytic action and tissue destruction (Dutra and Golllob, 2008).

**Figure 1.8.** Immunoregulatory model in chronic indeterminate and cardiac Chagas disease. Source: Dutra and Golllob, 2008.
1.6 Current treatment and management of Chagas disease

There are currently two chemotherapies available to treat *T. cruzi* infection: nifurtimox (Bayer, Leverkusen, Germany) and benznidazole (LAFEPE, Recife, Brazil), of which the latter is the first-line treatment with the superior safety profile (Viotti et al., 2009). In general, anti-trypanosomal treatment is strongly recommended for all acute, congenital and reactivated cases and all children and patients up to 18 years of age with chronic disease (Rassi Jr et al., 2010). The treatment regimens for children are benznidazole (5-10 mg/kg daily) in two or three divided doses for 60 days, or nifurtimox (15 mg/kg daily) in three divided doses for 60-90 days. For adults, daily treatment with 5 mg/kg benznidazole or 8-10 mg/kg nifurtimox is recommended for the same duration as children (Rassi Jr et al., 2010). Treatment is contraindicated during pregnancy and in patients with severe renal or hepatic insufficiency or advanced chagasic cardiomyopathy or megaoesophagus.

The major limitations associated with both drugs are significant side effects (ranging from mild hypersensitivity and polyneuropathy to Stevens-Johnson syndrome or toxic epidermal necrolysis) (Yun et al., 2009; Jackson et al., 2010b; Pinazo et al., 2010), prolonged treatment regimens, lack of paediatric formulations (Rassi Jr et al., 2009) and considerable supply problems in endemic countries (Burki, 2011). Both drugs have significant activity in the acute phase with over 80% success rates in treated patients (Pinto et al., 2009). Efficacy during the chronic phase is much lower and varies with geographical location, which may reflect different methods of treatment evaluation, incomplete treatment administration, variable drug susceptibility among distinct *T. cruzi* strains or characteristics of the host’s immune response (Andrade et al., 1992; Toledo et al., 2003; Urbina, 2010; Zingales et al., 2014). The benefits of benznidazole treatment to halt Chagas disease progression in adults aged 19-50 without advanced cardiomyopathy, is presently debated, and the subject of two ongoing multi-centre, double-blind, placebo-controlled clinical trials (BENEFIT and TRAENA) (Marin-Neto et al., 2009).

For symptomatic cardiac patients with sustained ventricular tachycardia and myocardial dysfunction, who are at high risk of death from arrhythmias, amiodarone treatment may be recommended (Scanavacca et al., 1990). Patients with refractory or unstable ventricular tachycardia can also be treated with implantable cardioverter defibrillators (ICDs) (Martinelli et al., 2012) or pacemakers (Clark et al., 2014). However, these options are often prohibitively expensive for local indigenous populations (Clark et al., 2014). Future prospects for Chagas disease management include the identification of biomarker predictors of progression, allowing treatment, surgical interventions etc., to be targeted to the subset of chronically-infected patients with the highest risk of morbidity and mortality (Requena-Méndez et al., 2013).

1.7 Prospective treatment of Chagas disease

Recent impetus from international agencies has made significant achievements in advancing the evaluation of novel chemical classes for Chagas disease (Buckner, 2011). Two phase II clinical trials assessing the activity of ergosterol biosynthesis inhibitors, posaconazole
(Molina et al., 2014) and rauconazole E1224 pro-drug (DNDi, 2013), were recently completed, demonstrating no significant improvement over the current gold standard benznidazole. Additional drugs in the research and development pipeline include flexinidazole (5-nitroimidazole) (Bahia et al., 2012; 2014), vinyl sulfone derivative K777 (cysteine protease inhibitor) and fenarimol (fungicide) (Keenan et al., 2013a; 2013b; 2012).

Alternate strategies also under investigation in mouse models are combination or sequential therapy with the aim of reducing drug dosage and/or duration of treatment (Benaim et al., 2006; Bustamante et al., 2013; Diniz et al., 2013).

1.8 Vector species, behaviours and ecologies
The triatomine bugs responsible for T. cruzi transmission represent a diverse taxonomic group encompassing at least 140 species belonging to six tribes and 19 genera (Lent and Wygodzinsky, 1979; Galvão, 2003). T. cruzi is principally transmitted by five domestic vector species, belonging to three genera: Triatoma infestans, Rhodnius prolixus, Panstrongylus megistus, Triatoma brasiliensis and Triatoma dimidiata (Figure 1.9 and 1.10).

Figure 1.9. Major triatome bug species responsible for T. cruzi transmission. Top row from left: adults of Triatoma infestans, Rhodnius prolixus and Panstrongylus megistus. Bottom row from left:
nymphs of *Triatoma infestans* and *Rhodnius prolixus* and *Rhodnius* egg. Source: L.A. Messenger and C. Whitehorn (LSHTM).

Triatomiine bugs are nocturnal, obligate haematophages, which require at least one blood meal to develop through each of five nymphal instars. All developmental stages and adults of both sexes can transmit *T. cruzi*. The probability that a triatomine is infected is directly proportional to number of blood meals taken, thus older instars and adults tend to have the highest infection rates. Paradoxically, stecorarian transmission of *T. cruzi* to humans is highly inefficient, estimated to occur every 900-4000 infected contacts (Nouvellet *et al.*, 2013).

**Figure 1.10.** Geographical distribution of the 14 most important vector species of Chagas disease. Source: Patterson and Guhl, 2010.
The triatomine bug species responsible for *T. cruzi* transmission are restricted to the Americas, with the exception of *Triatoma rubrofasciata*, which has a global tropicopolitan distribution in association with *Rattus rattus* (Patterson and Guel, 2010). In general, the major taxonomic groups are broadly correlated with distinct geographies and many display evolutionary adaptations to specific ecological niches (Gaunt and Miles, 2000) (section 1.13) (Figure 1.10).

*Rhodnius* species primarily inhabit *Attalea* palm trees (Gaunt and Miles, 2000) across the Amazon (Abad-Franch and Monteiro, 2007), the northern Andean countries, and parts of Central America. *Triatoma* species principally colonize terrestrial, rocky habitats, concentrated around southern and central Brazil and neighbouring countries, with a few species extending into the southern United States, where they have been implicated in sporadic autochthonous transmission of *T. cruzi* (Stevens et al., 2012; Garcia et al., 2014). Historically the maximum distribution of the principal domestic vector in South America, *Triatoma infestans*, ranged from the 12 most populated states of Brazil across vast areas of Bolivia, Argentina, southern Peru, northern Chile, Paraguay and Uruguay (Schofield et al., 2006). Nowadays, domestic *T. infestans* is confined to the Andean valleys of Bolivia and the Gran Chaco region in sympatry with several wild populations (Cortez et al., 2007; 2008; Buitrago et al., 2010; Ceballos et al., 2011; Waleckx et al., 2012).

*Panstrongylus* species, found predominantly in burrows and tree cavities, can be divided into two phylogenetically defined groups (Lent and Wygodzinsky, 1979; Marcilla et al., 2002) which correspond to regions north and south of the Amazon and East and West of the Andes (Patterson et al., 2009). *P. geniculatus* has the most extensive uninterrupted distribution of any triatomine species, ranging from the Atlantic to Pacific coasts and from mid-Central America to northern Argentina (Abad-Franch and Monteiro, 2007).

Dispersal of triatomines can be active or passive. Active transportation among domestic species is facilitated by walking, and less frequently by flying (Noireau and Dujardin, 2001); the average dispersal range of *T. infestans* is <500m (Richer et al., 2007). Passive dispersion involves transportation of usually immature stages by sylvatic vertebrate hosts (e.g. eggs glued to feathers (Forattini et al., 1971)) or attached to familiar objects carried or worn by humans and is the most important in the context of Chagas disease epidemiology (Noireau and Dujardin, 2010). Phylogenetic studies suggest that historically the main domestic populations of *T. infestans* were passively distributed by anthropogenic migration (Dujardin, 1998; Bargues et al., 2006; Piccinali et al., 2009; Cortez et al., 2010). This dependency on human hosts and domestic animals, overall lack of intra-species genetic diversity and loss of genetic resources from sylvatic populations (Panzera et al., 2004), would theoretically render this species highly susceptible to chemical control measures. These observations formed part of the rationale for the large trans-national vector control initiatives conducted during the 1990s (section 1.10). In practice, contemporary insecticide resistance threatens the success of local control programmes in areas of residual vector transmission, such as the Bolivian Chaco region (Germano et al., 2010; Lardeux et al., 2010).
1.9 Mammalian reservoir hosts and transmission cycle dynamics

More than 150 species of domestic (e.g., dogs, cats and guinea pigs), peridomestic (e.g., rodents, goats and pigs) and wild mammals (e.g., bats, marsupials and primates), from eight different orders, have been reported as infected with *T. cruzi*, although it is widely believed that all mammals are susceptible (Noireau et al., 2009a). Birds and cold-blooded vertebrates are refractory to infection (Kierszenbaum et al., 1981). While prevalence rates are not well established, certain species, particularly *Dasypus novemcinctus* (nine-banded armadillo), *Didelphis marsupialis* and *D. albiventris* (opossums), appear to be prominent sylvatic reservoirs of infection (Figure 1.11) (Yeo et al., 2005). *T. cruzi* is able to colonize almost all tissues in its mammalian hosts, including unconventional sites such as the cornea of *Thrichomys apereoides* (Herrera et al., 2007a) and the anal scent glands of *Didelphis* species (Deane et al., 1984a), enabling the latter species to function as both a host and vector of *T. cruzi*.

![Figure 1.11. Examples of important sylvatic reservoirs of *T. cruzi* infection: *Didelphis albiventris* (left) and *Dasypus novemcinctus* (right). Source: Noireau et al., 2009a.](image)

In addition to vector-borne infection, many sylvatic mammals are prone to alternate transmission routes, including oral infection via predation of infected vectors or mammals (Jansen and Roque, 2010) and exposure to contaminated anal scent gland secretions (Deane et al., 1984a; Carreira et al., 2001). These biological features may predispose such hosts to multiplicity of infection which is directly related to intensity and efficiency of parasite transmission and duration and course of disease (Roellig et al., 2010; Nouvellet et al., 2013). In terms of *T. cruzi* transmission, ‘maintenance reservoirs’ are considered to be those that are able to be infected and retain the infection, while ‘amplifier reservoirs’ are those that display characteristics of infection that favour parasite transmission, e.g. high parasitaemia levels (Jansen and Roque, 2010).
Chagas disease transmission cycles can be described as domestic, peridomestic or sylvatic and separate or overlapping in their ecology (Miles et al., 2009). Domestic transmission arises when triatomine vector species colonise human settlements, feeding exclusively on the inhabitants (Figure 1.12).

**Figure 1.12.** Example of domestic transmission in an endemic region of the Gran Chaco, Santa Cruz department, Bolivia. A: Gutierrez municipality, Cordillera province, where the rural population is almost exclusively of Guarani ethnicity and Chagas disease seroprevalence is 97% in adults older than 30 years. B: Typical house constructed from mud and sticks (tabique). C: *T. infestans* collected from inside B. D. Triatomine faeces stains, indicative of heavy house infestation. Source: L.A. Messenger (LSHTM).

Peridomestic foci, including chicken coops, goat corrals and pigsties, act as sites of triatomine re-invasion following the eradication of domestic populations (Cecere et al., 2006; Breniere et al., 2007) (Figure 1.13). Infection restricted to wild triatomines and reservoir hosts is termed sylvatic (or enzootic) transmission. Direct human infection is rare but sporadic cases can arise when adventitious sylvatic vectors, such as *T. sordida*, *R. pictipes* and *P. geniculatus*, fly into human settlements attracted by light (Miles et al., 1981b), *R. brethesi* attacks workers harvesting piassaba palms (Coura et al., 2002) or through oral
transmission when infected bugs enter juice presses, such as those used for açai palm or sugar cane (section 1.2).

**Figure 1.13.** Example of potential peridomestic reservoir hosts in an endemic region of the Gran Chaco, Santa Cruz department, Bolivia. A: Goats. B: Pigs. C: Dogs. D: Ducks and ducklings. All photos were taken either inside or less than 10 m from a house. Source: L.A. Messenger (LSHTM).

Separate transmission cycles are the most feasible to interrupt as there is limited potential for vector migration from the surrounding environment. However, in some epidemiological settings, domestic and sylvatic transmission cycles overlap, threatening the success of control programmes, specifically by the re-invasion of sylvatic vectors after the elimination of domestic colonies (Fitzpatrick *et al.*, 2008) and/or the adaptation of sylvatic species to inhabit the domestic or peridomestic niche (Matias *et al.*, 2003; Carrasco *et al.*, 2005; Noireau *et al.*, 2005; 2009b).

### 1.10 Control of Chagas disease

Considering the limited success of chemotherapy and absence of a prophylactic vaccine for Chagas disease (Vazquez-Chagoyan *et al.*, 2011), control strategies have traditionally concentrated on interrupting vector-borne transmission through the elimination of domiciliary triatomine species. Since 1990, case reporting of Chagas disease has decreased
by ~70%, following the success of coordinated multi-country vector control programmes, most notably the Southern Cone Initiative (INCOSUR) (Figure 1.14) (Moncayo, 2003).

In 1991, the governments of Argentina, Bolivia, Brazil, Chile, Paraguay and Uruguay, under the auspices of the Pan American Health Organization (PAHO; http://www.paho.org), initiated a trans-national control programme aimed at the total eradication of the principal domestic vector *T. infestans*, the suppression of secondary vector species and compulsory blood-bank screening (Schofield and Dias, 1999). The basic preventive strategies consisted of chemical control of vectors and screening of blood donors using serology, supported by health education, community participation, improved housing construction and epidemiological surveillance. As of 2006, Uruguay, Chile and Brazil have been declared free of active transmission, while a number of provinces in Argentina, Bolivia and Paraguay have also reported disease interruption (Moncayo and Silveira, 2009).

![Figure 1.14. The seroprevalence of Chagas disease and geographical distribution of control programmes. Data according to the PAHO, 2006. Source: Coura and Vinas, 2010.](image)

In light of these achievements, several other programmes, including the Andean Initiative (IPA, 1996) (Guhl, 2007), the Central American Initiative (IPCA, 1997) (Ponce, 2007) and the Amazon Initiative (AMCHA, 2004) (Aguilar *et al.*, 2007) have been implemented with the goal of interrupting Chagas disease transmission through the establishment of nationwide
blood screening programmes and vector control, with an emphasis on preventing domestic re-infestation from sylvatic foci, as vectors in these areas are not solely domiciliated.

Whilst the progress made over the last 20 years to control Chagas disease is indisputable, major challenges remain to ensure the sustainability and prioritization of these programmes in areas with very low residual disease prevalence and negligible house infestation rates (Schofield et al., 2006).

More recently, initiatives in countries of non-endemicity have also been introduced to control the transmission of Chagas disease by blood transfusion and transplantations of organs from infected migrant donors. For example, the USA, Spain and France have established measures to minimize transfusional risk through screening of blood donors and deferral strategies (Castro, 2009).

1.11 The origins of T. cruzi

Two theories have been proposed to explain the origins of T. cruzi (Stevens et al., 1999; Hamilton et al., 2012). Phylogenetic analysis of 18S rRNA sequences suggested that salivarian trypanosomes (members of the T. brucei clade, transmitted in saliva) diverged from stercorarian trypanosomes (members of the T. cruzi clade, transmitted in faeces) approximately 100 million years ago (MYA), which coincided with the breakup of Gondwana, the continental landmass that included Africa, South America, Antarctica and Australia (Stevens et al., 1999; Stevens and Gibson, 1999; Stevens et al., 2001) (“southern super-continent hypothesis”) (Figure 1.15a). Based on this hypothesis, considerable species diversity would be expected within the T. cruzi clade, assuming its members co-evolved with South American terrestrial mammals since the separation of South America from Antarctica and Australia ~40 MYA (Hamilton et al., 2012). However, no additional ‘bona fide’ T. cruzi clade species have been isolated from South American mammals, while T. cruzi clade trypanosomes have been identified in African terrestrial mammals, implying geographical isolation may not have been entirely responsible for the evolution of the T. cruzi clade (Hamilton et al., 2009).
Recent molecular analyses favour the premise that \textit{T. c. cruzi} evolved from a bat trypanosome ("bat seeding hypothesis"), given that it is most closely related to \textit{T. cruzi marinkellei} (from South American bats) (Hamilton et al., 2007; Cavazzana et al., 2010), \textit{Trypanosoma erneyi} (Lima et al., 2012), \textit{Trypanosoma livingstonei} (Lima et al., 2013) (both from bats in Mozambique) and \textit{Trypanosoma dioni}i (from Old and New World bats) (Hamilton et al., 2007; Hamilton et al., 2012), and is frequently isolated from bats (Lisboa et al., 2008; Marcili et al., 2009a; Pinto et al., 2012; Ramírez et al., 2014) (Hamilton et al., 2012) (Figure 1.15b). The “bat seeding hypothesis” proposes that various lineages of bat trypanosomes have independently switched hosts into terrestrial mammals, facilitated by invertebrate vectors feeding on both bats and mammals occupying the same ecological niche (Hamilton et al., 2012).

\textit{T. c. cruzi} has subsequently undergone major genetic diversification and dispersal across Central and South America. Current international consensus recognizes a minimum of six stable genetic lineages or discrete typing units (DTUs): Tc1-TcV1 (Zingales et al., 2009), which have distributions loosely defined by geography, ecology and transmission cycle (Miles et al., 2009) (section 1.13).
Contemporary population structure of T. cruzi

*T. cruzi* displays remarkable genetic diversity and this has long been implicated as one of the principal factors underlying the major geographical, biology, epidemiological and clinical variation observed in Chagas disease (Miles *et al.*, 1981a; Miles *et al.*, 2009). Molecular analyses indicate that *T. cruzi* has a predominantly clonal population structure, interspersed with infrequent genetic exchange events. DTUs TcI-TcIV form monophyletic clades and TcV and TcVI are known to be recent natural inter-lineage hybrids (Machado and Ayala, 2001; Lewis *et al.*, 2011). Multilocus sequence typing (MLST) supports these designations with TcI-TcIV characterized by substantial allelic homozygosity, likely resulting from recurrent, genome-wide and dispersed gene conversion, while TcV and TcVI display natural heterozygosity and minimal distinction, sharing intact alleles from their parental progenitors (TcII and TcIII) (Machado and Ayala, 2001; Brisse *et al.*, 2003; Barnabé *et al.*, 2011; Lewis *et al.*, 2011; Yeo *et al.*, 2011).

The origin(s) of these hybrid DTUs is unresolved and it is presently contested whether they arose from two independent genetic exchange events (Figure 1.16b) (de Freitas *et al.*, 2006; Lewis *et al.*, 2011) or a single incidence of hybridization followed by clonal divergence (Figure 1.16a) (Westenberger *et al.*, 2005; Sturm and Campbell, 2010; Flores-López and Machado, 2011). It has also been suggested that TcIII and TcIV are the products of a more ancient recombination event between TcI and TcII, which have subsequently undergone extensive loss of heterozygosity (LOH) (Figure 1.16a) (Westenberger *et al.*, 2005).

**Figure 1.16.** Comparison of the two-hybridization (a) and three ancestor (b) models which describe the role of genetic exchange during the evolution of *T. cruzi*. Rectangles represent the six *T. cruzi* DTUs. Genetic exchange is indicated by ovals, with parental contributions designated by red arrows. The three mitochondrial clades identified by de Freitas *et al.*, are highlighted in different colours: blue = clade A; green = clade B; orange = clade C. Source: Zingales *et al.*, 2012.
1.13 Molecular epidemiology of T. cruzi

The epidemiological relevance of the T. cruzi DTUs has also been the subject of considerable debate, with evidence emerging to support historical and contemporary associations of particular lineages with different transmission ecologies (Figure 1.17) (Table 1.1). In general, TcI, TcII, TcV and TcVI are frequently isolated from domestic cycles and are responsible for the majority of human infections. The distribution of domestic TcI extends from the Amazon Basin northwards, where it is the principal cause of Chagas disease in Venezuela and Colombia (Anez et al., 2004; Ramírez et al., 2010; Carrasco et al., 2012). TcI is also ubiquitous in arboreal sylvatic transmission cycles throughout Latin America (Barnabé et al., 2000; Roellig et al., 2008), primarily circulating in arboreal ecotopes between Didelphis species and the triatomine tribe Rhodniini (Gaunt and Miles, 2000), with secondary terrestrial cycles among rodents and sylvatic Triatoma species in the inter-Andean valleys of Bolivia, Peru and Chile (Cortez et al., 2006; Barnabé et al., 2011; Arenas et al., 2012; Breniere et al., 2012). Multiple molecular markers consistently identify high levels of genetic diversity within sylvatic TcI populations (Herrera et al., 2007b; 2009; O’Connor et al., 2007; Falla et al., 2009; Llewellyn et al., 2009a; Ocaña-Mayorga et al., 2010; Lima et al., 2014), and divergent, but genetically homogeneous, strains associated with human infections (Llewellyn et al., 2009a; Cura et al., 2010; Ramírez et al., 2012; Zumaya-Estrada et al., 2012).
Table 1.1. An overview of ecotopes, sylvatic vectors/hosts, geographical distributions and clinical associations of the major *T. cruzi* DTUs. Adapted from Miles *et al.*, 2009.

<table>
<thead>
<tr>
<th>DTU</th>
<th>Ecological Niche</th>
<th>Sylvatic Vectors</th>
<th>Sylvatic Hosts</th>
<th>Geographical Distribution</th>
<th>Clinical Forms of Human Chagas Disease</th>
</tr>
</thead>
<tbody>
<tr>
<td>TcI</td>
<td>Primary: lowland tropical arboreal</td>
<td>Primary: <em>Rhodnius</em> spp.</td>
<td>Primary: Arboreal marsupials (<em>Didelphis</em>, <em>Monodelphis</em>), primates, caviomorphs</td>
<td>Primary: southern United States, Central and South America</td>
<td>Cardiomyopathy</td>
</tr>
<tr>
<td>TcII</td>
<td>Rare in sylvatic cycles</td>
<td>ND</td>
<td>Atlantic forest primates</td>
<td>Atlantic/Central Brazil and Southern Cone</td>
<td>Cardiomyopathy GI megasyndromes Congenital infections</td>
</tr>
<tr>
<td>TcIII</td>
<td>Terrestrial, fossorial, lowland, arid and tropical</td>
<td><em>P. geniculatus</em></td>
<td>Armadillos (<em>D. novemcinctus</em>, <em>Euphractus sexcinctus</em>, <em>Chaetophractus</em>) marsupials (<em>Didelphis</em>, <em>Monodelphis</em>), rodents, carnvores</td>
<td>Northeastern Venezuela to Argentina</td>
<td>Rare in humans</td>
</tr>
<tr>
<td>TcIV</td>
<td>Arboreal with terrestrial transmission in North America</td>
<td>Rhodnius, <em>Panstrongylys</em>, <em>Triatoma</em></td>
<td>Primates, <em>D. novemcinctus</em>, <em>Nasua nasua</em>, <em>Procyon lotor</em></td>
<td>Southern United States and Northern South America</td>
<td>Secondary agent in Venezuela Sporadic oral outbreaks in Brazilian Amazon</td>
</tr>
<tr>
<td>TeV</td>
<td>Rare in sylvatic cycles</td>
<td>ND</td>
<td>ND [Canis familiaris]</td>
<td>Principally Southern Cone, Gran Chaco</td>
<td>Cardiomyopathy GI megasyndromes Congenital infections</td>
</tr>
<tr>
<td></td>
<td>Putative peridomestic transmission among dogs</td>
<td></td>
<td></td>
<td>Sporadic reports in Colombia</td>
<td></td>
</tr>
<tr>
<td>TcVI</td>
<td>Rare in sylvatic cycles</td>
<td>ND</td>
<td>ND [Canis familiaris]</td>
<td>Principally Southern Cone, Gran Chaco</td>
<td>Cardiomyopathy GI megasyndromes Congenital infections</td>
</tr>
<tr>
<td></td>
<td>Putative peridomestic transmission among dogs</td>
<td></td>
<td></td>
<td>Sporadic reports in Colombia</td>
<td></td>
</tr>
<tr>
<td>TeBat</td>
<td>Not described</td>
<td>ND</td>
<td><em>Chiroptera</em> spp.</td>
<td>Panama, Central and South East Brazil, Colombia</td>
<td>One isolated case of human infection reported</td>
</tr>
</tbody>
</table>
By contrast, TcII, TcV and TcVI are less genetically diverse overall (Lewis et al., 2011) and appear confined to domestic transmission cycles in southern parts of South America (Miles et al., 2009). The sylvatic reservoirs of these three DTUs are not fully defined, although TcII has been increasingly isolated from primates in Brazil (Fernandes et al., 1999; Lisboa et al., 2007; Araújo et al., 2011); peridomestic dogs are emerging as potential hosts of TcV and TcVI (Maffey et al., 2012; Enriquez et al., 2013; Fernandez et al., 2014). The geographical range of TcV and TcVI appears to be more extensive than previously suggested, with reports of these hybrid DTUs as far north as Colombia (Guhl and Ramírez, 2013; Messenger et al., submitted).

Figure 1.17. Approximate geographical distributions of T. cruzi lineages and transmission cycles. Source: M. Llewellyn (LSHTM).

TcIII has a dispersed terrestrial distribution that ranges from Amazonia to Argentina, where it is primarily transmitted by P. geniculatus to D. novemcinctus and other burrowing mammals (Yeo et al., 2005; Marcili et al., 2009b; Llewellyn et al., 2009b). TcIV is poorly understood, principally because several genotyping methods fail to distinguish this lineage from others, particularly from TcIII (Fernandes et al., 1998a; Lewis et al., 2009a). However, TcIV is known to circulate sympatrically with TcI in wild primates (Marcili et al., 2009c) and raccoons and dogs (Roellig et al., 2013), in the Amazon and North America, respectively. TcIV is also increasing in importance to human disease; it is a secondary agent of Chagas
disease in Venezuela (Miles et al., 1981a; Carrasco et al., 2012) and has been isolated from recent oral outbreaks in the Brazilian Amazon (Roque et al., 2008; Marcili et al., 2009c; Valente et al., 2009; Monteiro et al., 2010; 2012) and Colombia (Ramírez et al., 2013a). As yet, TcIII and TcIV only sporadically invade domestic transmission cycles, but this may reflect the relative paucity of available samples and the insensitivity of conventional genotyping methods. Finally, TcBat, a new genetically-distinct and potentially human-infective lineage (Ramírez et al., 2013b) has been isolated from Chiroptera species across Brazil (Marcili et al., 2009a), Panama (Pinto et al., 2012) and Colombia (Ramírez et al., 2014).

1.1.4 Genetic exchange in kinetoplastid protozoa and its phenotypic implications

Many eukaryotic pathogenic microorganisms (both fungal and protozoan) that were previously assumed to reproduce clonally have non-obligate cryptic sexual cycles (Heitman, 2010). Genetic exchange has the potential to drive the evolution of novel recombinant strains with epidemiologically significant traits, including increased pathogenicity, transmissibility and drug resistance (Awadalla, 2003). However, limiting sexual reproduction allows the generation of host-adapted clonal populations that retain the ability to hybridize in response to selective pressures (Heitman, 2006). The conservation of meiotic gene orthologues among several basally divergent protists, including Giardia (Poxleitner et al., 2008), Entamoeba (Ramesh et al., 2005) and Trichomonas (Malik et al., 2008) suggests that the common ancestor of all eukaryotes was capable of meiotic recombination.

Evidence of linkage disequilibrium (non-random association of genotypes at different loci) and deviations from Hardy-Weinberg expectations among various parasitic protozoa, including T. cruzi, T. brucei and Leishmania species, has reinforced the pervasive view that these pathogens rarely engage in sexual reproduction if even at all (Smith et al., 1993; Tibayrenc et al., 1986; 1990; reviewed by Tibayrenc and Ayala, 2012; 2013). However, the generation of in vitro experimental hybrids in each of these species indicates that all have retained their capacity for genetic exchange (Jenni et al., 1986; Gaunt et al., 2003; Akopyants et al., 2009; Calvo-Álvarez et al., 2014).

Molecular analyses of representative field isolates now suggest that a variety of population structures exists among natural geographical foci, ranging from endogamy/inbreeding (Rougeron et al., 2009; 2011) to epidemics (MacLeod et al., 2000) and exogamy/outbreeding (Morrison et al., 2009), which do not conform to traditional models of strict clonality or panmixia (Smith et al., 1993; Ramírez and Llewellyn, 2014) (Figure 1.18). The existence of natural hybrid strains has also been reported in T. cruzi (section 1.1.2) and between a number of different Leishmania sub-species (Belli et al., 1994; Dujardin et al., 1995; Ravel et al., 2006; Nolder et al., 2007; Rougeron et al., 2009; Chargui et al., 2009; Odiwuor et al., 2011; Gelanew et al., 2014; Rogers et al., 2014).
Figure 1.18. Mechanisms of genetic exchange among selected parasitic protozoa, predicted implications for population structure and observed population structures in the field. Source: Ramírez and Llewellyn, 2014.

In *T. brucei brucei* and *L. major*, genetic exchange is non-obligatory, occurs in their insect vectors (tsetse flies and sandflies, respectively) and involves Mendelian segregation of parental alleles (Jenni et al., 1986; Akopyants et al., 2009; Peacock et al., 2011; Inbar et al., 2013), and in the former, the production of putative haploid life cycle stages (Peacock et al., 2014). By comparison, characterization of experimental intra-TcI hybrids indicates that for *T. cruzi* recombination may arise within the mammalian host and is analogous to the parasexual cycle of *Candida albicans* (Bennett and Johnson, 2003); nuclear fusion creates a tetraploid intermediate, followed by homologous recombination, gradual genome erosion and reversion to aneuploidy (Gaunt et al., 2003; Lewis et al., 2009b). This unusual mating system differs from canonical meiosis and is challenging to reconcile with both the presence of highly conserved meiosis-specific orthologues within the *T. cruzi* genome and the existence of natural diploid heterozygous lineages, which resemble Mendelian F1 progeny (TcV and TcVI) (Lewis et al., 2009b; 2011). These hybrid strains are found almost exclusively in
domestic transmission cycles and are associated with human disease in Paraguay, Chile, Argentina, Bolivia and Southern Brazil, despite the abundance of other genotypes in these regions. Molecular dating indicates that these DTUs evolved recently, which suggests there is continuous risk of genetic exchange driving the emergence of novel recombinant genotypes (Lewis et al., 2011).

Importantly, the effect of hybridization on T. cruzi phenotype remains unresolved. The phenomenon of hybrid vigour (heterosis), whereby recombinants display novel, often enhanced phenotypes compared to parental strains, is well documented among parasitic protozoa (reviewed by Detwiler and Criscione, 2010). Preliminary reports of heterosis in T. cruzi, manifesting as different temperature stabilities of glucosephosphate isomerase isoenzymes, were described by Widmer et al., 1987. More recent observations of natural Leishmania hybrids indicate that genetic exchange can alter vector permissibility (Volf et al., 2007), increase virulence (Akopyants et al., 2009; Cortes et al., 2012) and generate successful variants which are capable of widespread clonal dissemination (Schwenkenbecher et al., 2006; Nolder et al., 2007), a scenario reminiscent of the dispersal of TcV and TcVI throughout the Southern Cone countries.

Detecting genetic exchange among natural T. cruzi populations is of profound importance considering the expansion of the hybrid lineages within the domestic niche, their association with clinical disease and the capacity for recombination to accelerate the evolution of novel strains with potentially important phenotypes. Furthermore, understanding the impact of hybridization on parasite genetic diversity and population structure is crucial to provide an accurate assessment of the epidemiological risk associated with recombinant T. cruzi genotypes.
1.15 PhD aims, biological objectives and milestones

The overall aim of this PhD project was to investigate genetic diversity in *T. cruzi* through significant development of phylogenetic markers and their application to the characterization of natural parasite population structures and genetic exchange mechanisms.

Biological objectives were:

- To contribute to the characterization of inter- and intra-DTU *T. cruzi* genetic diversity among natural parasite populations.
- To investigate intra-host *T. cruzi* multiclonality and its association with transmission and the clinical status of Chagas disease.
- To explore the interaction between parasite genetic heterogeneity and ecological biodiversity.
- To examine the frequency and mechanisms of genetic exchange among natural parasite populations.

Specific milestones included to:

1. Evaluate nuclear housekeeping genes as MLST candidates for *T. cruzi* DTU-level assignment.
2. Optimize nuclear targets for inclusion in a standardized MLST scheme for wider use by the *T. cruzi* research community.
3. Develop mitochondrial phylogenetic markers to describe intra-lineage genetic diversity in combination with multilocus microsatellite typing (MLMT).
4. Investigate intra-host parasite multiclonality through the application of Illumina amplicon deep sequencing markers.
5. Exploit high resolution nuclear and mitochondrial genotyping to identify ecological determinants of sylvatic parasite diversification and to detect incidences of natural genetic exchange occurring among putative parasite contact zones.
2. Materials and methods

2.1 Ethics statements
All molecular analyses in chapters 3-5 were performed at the London School of Hygiene and Tropical Medicine (LSHTM) following approval from the Ethics Committee (document N° 5483 “Comparative epidemiology of genetic lineages of Trypanosoma cruzi”). All T. cruzi field strains were collected with permissions from local institutional review boards (IRBs). Written informed consent was obtained prior to isolating human-derived strains, under the supervision and approval of local IRBs.

2.2 Parasite stocks
T. cruzi strains used in chapters 3-5 were obtained from the LSHTM cryobanks or through members of the ChagasEpiNet EC consortium. Specific panels of strains are included in each individual study.

2.3 Routine procedures

2.3.1 Parasite (epimastigote) culture
T. cruzi epimastigotes were cultivated in either biphasic or liquid growth medium. Biphasic cultures consisted of 4% (w/v) blood agar base, 0.6% (w/v) agar, 0.6% (w/v) NaCl, 0.5% (w/v) tryptone (all Sigma-Aldrich, UK) and 10% (v/v) defibrinated rabbit blood, prepared in sterile Nunclon™ Δ flat sided culture tubes (Nunc, Denmark) and overlaid with 0.9% NaCl (w/v) solution and 50µg/ml gentamycin (Sigma-Aldrich, UK). Liquid growth medium consisted of either RPMI-1640 (#R0883) or liver infusion tryptose (LIT). RPMI-1640 was supplemented with 0.5% (w/v) tryptone, 20mM HEPES buffer (pH 7.2), 30mM haemin, 10% (v/v) heat-inactivated fetal bovine serum (FBS), 2mM sodium glutamate, 2mM sodium pyruvate, 250µg/ml streptomycine, 250U/ml penicillin and 25µg/ml gentamycin (all Sigma-Aldrich, UK). LIT was prepared by dissolving 25g liver infusion broth (Difco™, Becton Dickinson, USA), 5g tryptone, 4g NaCl, 2g glucose, 0.4g KCl and 3.15g Na₂HPO₄ in 900 ml ddH₂O and adjusting the pH to 7.4 before adding 25g haemin dissolved in 1 ml 1N NaOH, 100 ml heat inactivated FBS, 250µg/ml streptomycin, 250U/ml penicillin and 25µg/ml gentamycin (all Sigma-Aldrich, UK).

Cultures in liquid medium were maintained in 25cm² flasks (Becton Dickinson, USA) and incubated at 28°C in a humidified atmosphere of 5% CO₂. Parasites were routinely maintained in logarithmic growth phase by re-passaging cultures at 1:20-1:50 dilutions every 1-3 weeks, depending on strain growth rate.

For long-term cryopreservation of parasites, late logarithmic phase cultures were supplemented with sterile 10% glycerol (v/v) (VWR, UK) and aliquots were prepared in
cryovials (Nunc, Denmark). Cryovials were stored at -70°C for 24 hours before transfer to liquid nitrogen containers in the LSHTM cryobank.

To recover parasites from cryopreservation, cryovials were carefully thawed in a 37°C water bath and the contents aliquoted into a biphasic culture tube or 1 ml of liquid medium in a 24-well culture plate (Becton Dickinson, USA).

### 2.3.2 Isolating parasite field strains
*T. cruzi* strains were isolated from Chagas seropositive patients or mammalian hosts by hemoculturing; isolation protocols had minor differences between studies. In general, between 0.5-20 ml venous blood was collected in a sterile BD Vacutainer® spray-coated K$_2$ EDTA tube (Becton Dickinson, USA) and either aliquoted directly into biphasic culture tubes or LIT medium, or inoculated following the removal of plasma by centrifugation. All procedures were performed under sterile conditions in a laminar flow hood and samples were processed less than 12 hours after blood draw.

Hemocultures were maintained at 26-28°C and examined microscopically for motile trypanosomes fortnightly for up to six months. Positive primary cultures were allowed to grow to late logarithmic phase and preserved in 1:1 6M guanidine hydrochloride - 0.2M EDTA (GEB) (Sigma-Aldrich, UK).

### 2.3.3 Biological cloning of parasites on solid media
Solid phase cloning was undertaken at the LSHTM as described by Yeo *et al.*, 2007. Uncloned culture density was measured using a disposable haemocytometer (Immune Systems, UK) and $10^2$-$10^3$ logarithmic phase cells were mixed with 2.4 ml (w/v) supplemented RPMI-1640 medium and 0.6 ml molten 3% (w/v) low melting point (LMP) agarose containing 0.9% NaCl (w/v) and 50µg/ml gentamycin (all Sigma-Aldrich, UK). This overlay was poured onto a blood agar plate (prepared by adding 10.8 ml biphasic medium, described in section 2.3.1, to a sterile 90mm petri dish (Sterilin, UK)) and allowed to set. Plates were sealed with parafilm (VWR, UK) and incubated at 28°C in a humidified atmosphere of 5% CO$_2$.

Once colonies became visible (2-4 weeks depending on strain growth rate), they were examined microscopically, picked using a 200µl sterile pipette tip and inoculated into 1 ml liquid growth medium in a 48-well culture plate (Becton Dickinson, USA).

For studies in chapter 5, where parasite DNA was provided by Universidad de los Andes, biological clones were derived from primary parasite cultures by either limiting dilution (Ramírez *et al.*, 2013c) or fluorescence activated cell sorting (FACS) (Valadares *et al.*, 2012).
2.3.4 Extraction of parasite genomic DNA
Total genomic DNA from late logarithmic phase cultures (~10^7-10^8 parasites) was prepared using a Gentra Puregene Tissue kit (Qiagen, UK), according to the manufacturer’s protocol. For field samples in GEB, DNA was extracted using either a DNeasy® Blood and Tissue kit (QIAGEN, UK) or a High Pure PCR template preparation kit (Roche, UK), according to the manufacturer’s protocol.

2.3.5 Polymerase chain reaction (PCR)
Unless otherwise specified, all PCR reactions contained: 1X NH₄ reaction buffer, 1.5mM MgCl₂ (Bioline, UK), 0.2mM dNTPs (New England Biolabs, UK), 10pmol of each primer, 1U BIOTAQ™ DNA polymerase (Bioline, UK) and 10-100ng of template DNA, diluted in ddH₂O to a final volume of 20µl. Depending on sample numbers, reactions were prepared in either 0.2 ml PCR tube strips (VWR, UK) or 96-well reaction plates (Fisher Scientific, UK). All PCR amplifications were performed in a G-Storm thermal cycler (G-Storm, UK).

2.3.6 Agarose gel electrophoresis
PCR products were routinely visualized by gel electrophoresis using agarose gels containing 0.5µg/ml ethidium bromide (EtBr) (Sigma-Aldrich, UK). For DNA fragments with sizes ranging from 0.5-10kb, 1.5% agarose gels (Bioline, UK) were used and run at 90V for 1-4 hours in 1X TAE buffer, depending on the level of separation required.

In some cases, to visualize small base pair differences (5-20 bp), PCR products were separated by gel electrophoresis using 3.5% NuSieve™ GTG™ low melting temperature agarose gels (Lonza, UK) containing 0.5µg/ml ethidium bromide (Sigma-Aldrich, UK).

If PCR products required gel extraction, products were separated in 0.8% agarose gels (Bioline, UK). The following DNA molecular weight markers were used according to the range of size separation required: 0.1-1kb Hyperladder™ IV; 0.2-10kb Hyperladder™ I (both Bioline, UK). All samples were loaded into agarose gel wells with 1µl of 5X DNA loading buffer (Bioline, UK). Following electrophoresis, gels were visualized using a UV transilluminator (UVP).

2.3.7 PCR purification
All PCR products were purified using QIAquick PCR Purification kits (Qiagen, UK), according to the manufacturer’s protocol. When reactions produced multiple bands or a specific band was required for further analysis, the appropriate-sized band was excised with a sterile scalpel blade (SLS, UK) under UV illumination, and purified using a QIAquick Gel Extraction kit (Qiagen, UK).
2.3.8 Chain termination DNA sequencing

Bi-directional sequencing was performed using a BigDye™ Terminator v3.1 Cycle Sequencing kit (Applied Biosystems, UK), according to a modified version of the manufacturer’s protocol. Sequencing reactions contained 0.5µl Big Dye sequencing RR-100, 1.7µl sequencing buffer, 3.2pmol of forward or reverse primer and 20-100ng of purified PCR product or plasmid DNA, diluted in ddH2O to a final volume of 10µl. Reactions were subjected to 25 cycles of the following conditions: rapid thermal ramp to 96°C (1°C/second), 96°C for 30 seconds, rapid thermal ramp to 55°C (1°C/second), 55°C for 20 seconds, rapid thermal ramp to 60°C (1°C/second) and 60°C for 4 minutes.

Reactions were transferred to 96-well optical reaction plates with barcodes (Applied Biosystems, UK). For each sample, DNA was precipitated by the addition of 8µl ddH2O followed by 32µl ice-cold 95% (v/v) ethanol. Samples were incubated at 4°C for 15 minutes and then centrifuged at 3000g for 45 minutes at 4°C. Supernatants were removed by inverting the plates onto absorbent paper and centrifuging at 20g for 10 seconds. DNA pellets were washed by the addition of 50µl ice-cold 70% (v/v) ethanol and briefly vortexing. Samples were then centrifuged at 3000g for 30 minutes at 4°C. Supernatants were aspirated as previously and the DNA pellets left to dry at room temperature until no visible ethanol remained. DNA pellets were then re-suspended in 10µl Hi-Di™ deionized formamide (Applied Biosystems, UK) and stored at -20°C. DNA sequences were analyzed on a 16-capillary 3730 DNA Analyzer (Applied Biosystems, UK).

2.3.9 Ligation reactions

All ligation reactions were performed using reagents from the pGEM®-T Easy Vector System I kit (Promega, UK), according to the manufacturer’s instructions. Since TA cloning is dependent on the presence of a 3’terminal adenine base, added to amplicons during PCR amplification, all PCR products were first subjected to A-tailing by incubation for 30 minutes at 72°C in a reaction containing 1X NH4 reaction buffer, 2mM MgCl2, 10mM dATP and 2U BIOTAQ™ DNA polymerase (all Bioline, UK) diluted in ddH2O to a final volume of 50µl. A-tailed PCR products were then gel purified (section 2.3.7) and cloned separately into the pGEM®-T Easy Vector (Promega, UK).

Ligation reactions contained insert and vector DNA in a 3:1 molar ratio, 2X rapid ligation buffer, 50ng pGEM®-T Easy Vector and 3U T4 DNA ligase, diluted in ddH2O to a final volume of 10µl. Each set of reactions was accompanied by a self-ligation control (pGEM®-T Easy Vector with no insert) and two positive control reaction (control insert DNA with pGEM®-T Easy Vector or transformed whole plasmid DNA). Ligations were incubated at room temperature overnight.

2.3.10 Preparation of chemically competent bacterial cells

Bacterial cells from a frozen glycerol stock were streaked onto a lysogeny broth (LB) agar plate with no antibiotic selection, using a sterile 200µl pipette tip, and incubated overnight at
A single colony from this plate was inoculated into 2 ml LB broth with no antibiotic selection in a sterile 30 ml universal tube (Sterilin, UK) and the culture was grown overnight in a shaking incubator at 37°C. The following day 1 ml of the overnight culture was used to seed 100 ml of LB broth (no antibiotic selection) in a 2 L conical flask (Fisher, UK). The culture was grown in a shaking incubator at 37°C until the optical density at 600nm (measured using a spectrophotometer and calibrated using a cuvette of LB) reached between 0.6-0.8. Cells were transferred to a 50 ml centrifuge tube (Corning, UK) on ice and left to cool for 1 hour. The culture was centrifuged at 3500 rpm for 15 minutes at 4°C, the supernatant discarded, and the pellet re-suspended in 30 ml ice-cold 100mM CaCl₂ (containing 15% glycerol, VWR, UK) and incubated on ice for 1 hour. The cell suspension was centrifuged as previously and the pellet re-suspended in 8 ml ice-cold 100mM CaCl₂ (15% glycerol) and incubated on ice for 1 hour. Competent cells were aliquoted into sterile 1.5 ml graduated microtubes (Anachem, UK) and flash-frozen by incubating in dry ice and absolute ethanol for 5 minutes. Aliquots were stored at -80°C.

2.3.11 Bacterial transformations

Completed ligation reactions were transformed into Escherichia coli strain XL1-Blue (Agilent Technologies, UK), according to the following procedure. Sterile 1.5 ml graduated microtubes (Anachem, UK) were placed on ice and allowed to cool. To each tube 100µl of bacterial cell suspension was added, mixed with 5µl of ligation reaction, and incubated on ice for 30 minutes. Each tube was then heat shocked by incubating in a water bath at 42°C for 1 minute after which it was returned directly to ice and incubated for a further 2 minutes. 250µl of room temperature super optimal broth with catabolite repression (SOC) medium was added to each tube and reactions were incubated for 1-1.5 hours at 37°C in a shaking incubator. 100µl of each reaction was then spread aseptically onto the surface of a pre-warmed LB agar plate supplemented with 100µg/ml ampicillin (Sigma-Aldrich, UK), 80µg/ml 5-bromo-4-chloro-3-indolyl-β-D-galactopyranoside (X-gal) and 20mM isopropyl β-D-1-thiogalactopyranoside (IPTG) (both Bioline, UK). Agar plates were incubated at 37°C for 16-24 hours and stored at 4°C to facilitate colour-based colony selection.

Colonies were picked from agar plates using sterile 200µl pipette tips, which were then used to inoculate 5 ml cultures of fresh LB broth containing 100µg/ml ampicillin in sterile 30 ml universal tubes (Sterilin, UK). Cultures were incubated overnight in a shaking incubator at 37°C.

Plasmid DNA was purified from 3 ml of each overnight culture using a QIAprep Spin Miniprep kit (Qiagen, UK), according to the manufacturer’s protocol. Successful ligation of inserts into the pGEM®-T Easy Vector was confirmed by digestion of plasmid DNA with either EcoRI or NotI restriction enzymes, as described in section 2.3.12.
2.3.12 Restriction enzyme digests
Restriction enzyme digestion reactions typically contained up to 100ng/µl DNA, 0.25U/µl of one or two restriction endonucleases, 100ng/µl bovine serum albumin (BSA) and 1X manufacturer’s recommended reaction buffer diluted in ddH2O (all New England Biolabs, UK). All reactions were incubated at 37ºC for up to 24 hours and then visualized by agarose gel electrophoresis, as described in section 2.3.6.

2.4 T. cruzi genotyping

2.4.1 T. cruzi DTU assignment
Strains were routinely assigned to DTU-level using a standardized triple-assay comprising PCR product size polymorphism analysis of the 24Sα rRNA gene (LSU rDNA) and PCR-restriction fragment-length polymorphism analysis (PCR-RFLP) using heat shock protein 60 (HSP60) and glucose-6-phosphate isomerase (GPI) (Lewis et al., 2009a).

PCR amplifications were performed as described in section 2.3.5, using primers listed in Table 2.1. Reaction conditions for the 24Sα rRNA (LSU rDNA) PCR were an initial denaturation step of 94ºC for 3 minutes and then 27 amplification cycles (94ºC for 1 minute, 60ºC for 1 minute, 72ºC for 1 minute), followed by a final elongation step at 72ºC for 5 minutes. Reaction conditions for both HSP60 and GPI used a touchdown PCR strategy comprising an initial denaturation step of 3 minutes at 94ºC, followed by four cycles (94ºC for 30 seconds, 64ºC for 30 seconds, 72ºC for 1 minute), followed by 28 cycles (94ºC for 30 seconds, 60ºC for 30 seconds, 72ºC for 1 minute), and then a final elongation step at 72ºC for 10 minutes.

<table>
<thead>
<tr>
<th>PCR-RFLP Target</th>
<th>Primer Name</th>
<th>Primer Sequence (5’→3’)</th>
</tr>
</thead>
<tbody>
<tr>
<td>LSU rDNA⁵</td>
<td>D71</td>
<td>AAGGTGCGTGACAGTGTGG (20)</td>
</tr>
<tr>
<td>D72</td>
<td>TTTTCAGAATGGCCGAACAGT (21)</td>
<td></td>
</tr>
<tr>
<td>HSP60⁶</td>
<td>HSP60_for</td>
<td>GTGGTATGGGTGACATGTAC (20)</td>
</tr>
<tr>
<td>HSP60_rev</td>
<td>CGAGCAGCAGCGAACACGT (21)</td>
<td></td>
</tr>
<tr>
<td>GPI</td>
<td>GPI_for</td>
<td>GCCATGTAAGCTTTTGGGCTTCACATG (29)</td>
</tr>
<tr>
<td>GPI_rev</td>
<td>TGTAAGGCCCAGTGAGCGTGCCTGTTGAAATAGC (34)</td>
<td></td>
</tr>
</tbody>
</table>

⁵ Primer sequences according to Brisse et al., 2001.
⁶ Primer sequences according to Strurm et al., 2003.
⁷ Primer sequences according to Gaunt et al., 2003.
Successful amplifications were confirmed by visualization of 24Sα rRNA PCR products on 3.5% NuSieve™ GTG™ agarose gels (Lonza, UK) and HSP60 and GPI PCR products on 1.5% agarose gels (Bioline, UK), as described in section 2.3.6.

Table 2.2. T. cruzi genotype assignment of PCR amplification product sizes (bp).

<table>
<thead>
<tr>
<th>Target/Enzyme</th>
<th>Expected PCR product (digestion product) band size (bp)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>TcI</td>
</tr>
<tr>
<td>LSU rDNA</td>
<td>110</td>
</tr>
<tr>
<td>HSP60/EcoRV</td>
<td>432 - 462 (432 - 462)</td>
</tr>
<tr>
<td>GPI/HhaI</td>
<td>1,264 (817 + 447 + 253)</td>
</tr>
</tbody>
</table>

*According to Kawashita et al., 2001.
** Double band pattern observed for most isolates; 125bp band exhibits variable intensity
§According to Brisse et al., 2001.
¶For strains of North American origin, according to Brisse et al., 2001.

HSP60 and GPI PCR products were purified according to section 2.3.7 and then subjected to restriction digest with EcoRV or HhaI restriction endonucleases (New England Biolabs, UK), respectively, as described in section 2.3.12. 10µl of each digest reaction was visualized using either 1.5% (GPI/HhaI) or 3% agarose gels (HSP60/EcoRV) (Bioline, UK), as described in section 2.3.6.

T. cruzi strains were assigned to DTU-level based on the number and size of their restriction fragment bands, as shown in Table 2.2, Figure 2.1 and 2.2.
<table>
<thead>
<tr>
<th>LSU rDNA PCR product</th>
<th>110 bp only</th>
<th>125 bp only</th>
<th>110 bp only</th>
<th>120 bp*</th>
<th>110 bp or 110 + 125 bp</th>
<th>125 bp only</th>
</tr>
</thead>
<tbody>
<tr>
<td>HSP60/EcoRV PCR-RFLP</td>
<td>1 band</td>
<td>1 band</td>
<td>2 bands</td>
<td>1 band</td>
<td>3 bands</td>
<td>3 bands</td>
</tr>
<tr>
<td>GPI-Hhol PCR-RFLP</td>
<td>2 bands</td>
<td>3 bands</td>
<td>2 bands</td>
<td>3 bands**</td>
<td>4 bands</td>
<td>4 bands</td>
</tr>
</tbody>
</table>

*%I(IV/NA) = 150 bp
**%I(IV/NA) = 2 bands

**Figure 2.1.** Recommended triple-assay for discriminating *T. cruzi* DTUs. Source: Lewis et al., 2009a.
Figure 2.2. Examples of PCR-RFLP genotyping profiles. 
A. LSU rDNA. B. HSP60/EcoRV digestion products are shown. C. GPI/HhaI digestion products are shown. For all gels, Lanes: 1, Sylvio X10/1 (TcI); 2, Esm cl3 (TcII); 3, M5631 (TcIII); 4, CanIII cl1 (TcIV); 5, 92122102R (TcIV NA); 6, Sc43 cl1(TcV); 7, CL Brener (TcVI); 8, negative control.

2.4.2 Nuclear multilocus sequence typing (nMLST)
A total of nine single-copy nuclear housekeeping gene fragments were chosen for initial MLST evaluation: dihydrofolate reductase-thymidylate synthase (DHFR-TS), LYT1, metacyclin-II (Met-II), metacyclin-III (Met-III), RNA-binding protein-19 (RB19), ascorbate-dependent haemoperoxidase (TcAPX), glutathione-dependent peroxidase II (TcGPXII), mitochondrial peroxidase (TcMPX) and trypanothione reductase (TR). Primer sequences are given in Table 2.3. PCR reactions were performed as described in section 2.3.5.

For DHFR-TS and TR, reaction conditions were an initial denaturation step for 3 minutes at 94°C, followed by 30 amplification cycles (94°C for 30 seconds, 58°C for 1 minute, 72°C for 2 minutes). Annealing temperatures were 55°C for TcAPX, 50°C for TcMPX and 62°C for TcGPXII. Reaction conditions for Met-II and RB19 were 3 minutes at 94°C, followed by 30 amplification cycles (94°C for 30 seconds, 53°C for 30 seconds, 72°C for 45 seconds). Annealing temperatures were 51°C for Met-III and 56°C for LYT1. All reactions had a final ten minute elongation step at 72°C.

Successful amplifications were confirmed by visualization of PCR products on 1.5% agarose gels, as described in section 2.3.6. PCR products were purified according to section 2.3.7 and sequenced according to section 2.3.8 using PCR primer pairs and additional internal primers, as indicated in Table 2.3.

For optimization of MLST targets, PCR reactions contained: 10µl GoTaq® reaction buffer (Promega, UK), 0.2mM dNTPs (New England Biolabs, UK), 10pmol of each primer (Table
2.3), 1U GoTaq® DNA polymerase (Promega, UK) and 100ng of template DNA, diluted in ddH₂O to a final volume of 50µl.

Amplification conditions for all targets were an initial denaturation step for 5 minutes at 94°C, followed by 35 amplification cycles (94°C for 1 minute, 55°C for 1 minute, 72°C for 1 minute) with a final extension step of 72°C for 5 minutes.

Successful amplifications were confirmed by visualization of PCR products on 1.5% agarose gels, as described in section 2.3.6. PCR products were purified according to section 2.3.7 and sequenced according to section 2.3.8 using PCR primer pairs (Table 2.3). To assess reproducibility, each PCR amplification and subsequent sequencing reaction was repeated at least twice.

Additional MLST targets from Diosque et al., 2014, were used in selected analyses in chapter 5, as indicated.
### Table 2.3. *T. cruzi* nuclear gene fragments and primer details.

<table>
<thead>
<tr>
<th>Gene Fragment</th>
<th>Gene ID</th>
<th>Genome Position</th>
<th>Primer Sequence (5’→3’)</th>
<th>Amplicon Size (bp)</th>
<th>Sequence Start 5’</th>
<th>Sequence Start 3’</th>
<th>Fragment Length (bp)</th>
</tr>
</thead>
</table>
| *DHFR-TS*     | Tc00.1047     | 05350915        | 27: CGCTGTATTAAGATCCGNATGCC (22) FWD  
                            |                    | 27: CGCATAGTCAATGACCTCCATGTG (24) REV  
                            |                    | 27: ACCCTGTCCGTCAATAGTTG (19) INT’ | 1473 | GCGGGAGA | CACGCTCT | 715 |
| *LYT1*        | Tc00.1047     | 05350804        | 22: CAACTTGCCCTTTTGCTCTC (20) FWD  
                            |                    | 22: GTTGTGGTTGTGCCCTGTG (20) REV  
                            |                    | 22: ACAAGGGTGTTTCCGTAAG (20) INT | 804 | TTTGTGAC | GTGGTGGA | 691 |
| *Met-II*      | Tc00.1047     | 05351088        | 6: TCATCTGCAACCGATGACTTC (20) FWD  
                            |                    | 6: CTCCATAGCGTGCAAGCAACA (20) REV  
                            |                    | 6: TGTGTCCACAATTCGTCGT (20) INT | 702 | CATTTTCA | TTTTGCCA | 390 |
| *Met-III*     | Tc00.1047     | 05351094        | 36: GTGGCTCCAAAGGCATAGAAG (20) FWD  
                            |                    | 36: CCCCTTCTTTGCCAATTTAT (20) REV  
                            |                    | 36: CACATGAGTCGTTGTG (20) INT | 824 | TTCATCCG | TTTCTTG | 619 |
| *RB19*        | Tc00.1047     | 05350751        | 29: GCCTACACCGAGGAGTACCA (20) FWD  
                            |                    | 29: TTCTCCAATCCCAGACTTTG (20) REV | 408 | GTCGTGCCG | CCCAGACT | 350 |
| *TcAPX*       | Tc00.1047     | 05350619        | 36: GGGAGATTTTCCACCTTTGTGCTTTTCTGTAC (30) FWD  
<pre><code>                        |                    | 36: GGGGGATCCATGCCCTTTTGCTTTTGTTCA (30) REV | 1500 | GCGAGTCG | GGCGCCGG | 799 |
</code></pre>
<table>
<thead>
<tr>
<th>Gene</th>
<th>Accession</th>
<th>ID</th>
<th>q-value</th>
<th>Primers (FWD)</th>
<th>Precise Primers</th>
<th>Primers (REV)</th>
<th>Precise Primers</th>
</tr>
</thead>
<tbody>
<tr>
<td>TcGPXII</td>
<td>Tc00.1047 05351101</td>
<td>9.99</td>
<td>35</td>
<td>GGGCAGCAGCACGCTGTTCG (20)</td>
<td>TCATGCACCCCCGTTGCGGCCC (21)</td>
<td>TTGCGGCC ATCCGGCC</td>
<td>437</td>
</tr>
<tr>
<td>TcMPX</td>
<td>Tc00.1047 05350949</td>
<td>9.14</td>
<td>22</td>
<td>ATGTTTCGTCGTATGGCC (18)</td>
<td>TGCGTTTTTCTCAAATATTC (21)</td>
<td>TACATGGA CGACCGT</td>
<td>505</td>
</tr>
<tr>
<td>TR</td>
<td>Tc00.1047 05350355</td>
<td>5.30</td>
<td>37</td>
<td>ACTGGAGGCTGCTTGGAACGC (21)</td>
<td>GGATGCACACACRATRGTTGT (22)</td>
<td>TGTCAATG TACGAAGG</td>
<td>602</td>
</tr>
</tbody>
</table>

*Chromosomal assignment based on Weatherly *et al.*, 2009.

* Indicates internal primers used during sequencing reactions.
2.4.3 Phylogenetic analysis of nMLST

Nucleotide sequences were assembled manually in BioEdit v7.1.3.0 sequence alignment editor software (Ibis Biosciences, USA) (Hall, 1999) and unambiguous consensus sequences were produced for each isolate. Heterozygous SNPs were visually identified by the presence of two coincident peaks at the same locus (‘split peaks’), verified in forward and reverse sequences, and manually scored according to the one-letter nomenclature for nucleotides from the International Union of Pure and Applied Chemistry (IUPAC) (Figure 2.3).

![Sequence chromatogram](Image)

**Figure 2.3.** Example of a heterozygous sequence chromatogram. Four heterozygous SNPs in the coding region of Met-III from Sc43 cl1 (TcV) (heterozygous SNPs are indicated by arrows). From left to right the heterozygotes are C/G (=S), A/G (=R), A/G (=R) and C/T (=Y). This chromatogram emphasizes the importance of manually editing sequences as the first two split peaks have been incorrectly resolved by BioEdit and the latter two are ambiguous to the software.

Initially, phylogenetic analysis was performed on modified sequences according to the methodology described by Tavanti *et al.*, 2005, which renders diplotypic data suitable for the construction of distance-based dendrograms. Briefly, variable loci between isolates can be considered either homozygous or heterozygous, assuming that *T. cruzi* is minimally diploid. For each of the nine MLST targets, all SNPs were conjoined to derive a sequence type (ST) and each base in the sequence was duplicated, e.g. for homozygous variable loci (A, C, G or T) an “A” was rewritten as “AA”, while in the case of heterozygous loci, “Y” was scored as “CT” in accordance with IUPAC nomenclature (K=GT, M=AC, R=AG, S=CG, W=AT, Y=CT).

ST data were used to assess the discriminatory power (number of STs identified per total number of isolates) and typing efficiency (number of STs identified per polymorphic site) of each gene fragment (Hunter, 1990). For each target, the ratio of non-synonymous to synonymous amino acid changes (dN/dS) was calculated according to the Nei-Gojobori method (Nei and Gojobori, 1986), using SNAP software (available at [http://www.hiv.lanl.gov](http://www.hiv.lanl.gov)) (Korber, 2000) to infer relative selection pressures.

Modified sequences were used to generate neighbour-joining (NJ) dendrograms in MEGA 5.10 (Tamura *et al.*, 2011), to assess the “usefulness” of individual gene fragments for lineage assignment, intra-lineage resolution and to reconstruct phylogenetic relationships. Finally, variable loci data for all nine genes were concatenated together to produce a diploid
sequence type (DST) for each strain and a NJ phylogeny to evaluate the overall resolutive power of the MLST scheme.

2.4.4 Haplotype resolution
To infer haplotypes for each gene in selected analyses, diploid sequence data were submitted to PHASE software version 2.1 (Stephens et al., 2001) implemented in DnaSP v5.10.1 (Librado and Rozas, 2009). This program is based on a modified Markov chain-Monte Carlo (MCMC) algorithm which first assembles all unambiguous haplotypes, i.e. those observed in strains which are homozygous at all variable sites or heterozygous at only a single variable site. Haplotypes in the remaining isolates, which are heterozygous at multiple sites, (and therefore of ambiguous phase) are estimated using a Maximum-Likelihood (ML) approach, with the assumption that they are most likely to be either the same as, or closely related to, one of the known unambiguous haplotypes already sampled in the dataset. For each reconstructed haplotype, an estimation of the uncertainty associated with each phase call is also generated.

For isolates, where PHASE was unable to adequately resolve haplotypes (uncertainty probability \( p < 0.95 \)), PCR products were cloned and sequenced to experimentally verify predicted gene phase. PCR products were cloned using the pGEM®-T Easy Vector system I (Promega, UK), according to section 2.3.9. Plasmids were isolated from a minimum of six bacterial colonies, according to section 2.3.11, and sequenced, according to section 2.3.8, using standard T Easy Vector primers T7 (5’-TAATACGACTCACTATAGGG-3’) and Sp6 (5’-ATTTAGGTGACACTATAG-3’) (Promega, UK). In cases where haplotypes remained ambiguous a further six colonies were picked and processed, as described.

2.4.5 Maxicircle multilocus sequence typing (mtMLST)
Ten maxicircle gene fragments from eight coding regions were chosen for evaluation: \( ND4 \) (NADH dehydrogenase subunit 4), \( ND1 \) (NADH dehydrogenase subunit 1), \( COII \) (cytochrome c oxidase subunit II), \( MURF1 \) (maxicircle unidentified reading frame 1, two fragments), \( CYT b \) (cytochrome b), \( 12S rRNA \), \( 9S rRNA \) and \( ND5 \) (NADH dehydrogenase subunit 5, two fragments). An additional target (\( ND8 \); NADH dehydrogenase subunit 8) was also assessed and discarded due to poor PCR amplification. Degenerate primers were designed in primacleade (Gadberry et al., 2005) by reference to the complete maxicircle genomes from Sylvio X10/1 (TcI), Esm cl3 (TcII) and CL Brener (TcVI), available online at www.tritrypdb.org (Aslett et al., 2010). Primer sequences are given in Table 2.4.

PCR amplification for all targets was performed as described in section 2.3.5. Reaction conditions for all targets were an initial denaturation step of 3 minutes at 94°C, followed by 30 amplification cycles (94°C for 30 seconds, 50°C for 30 seconds, 72°C for 30 seconds) and a final elongation step at 72°C for ten minutes. Successful amplification was confirmed by
visualization of PCR products on 1.5% agarose gels, as described in section 2.3.6. PCR products were purified according to section 2.3.7 and sequenced according to section 2.3.8.
Table 2.4. *T. cruzi* maxicircle gene fragments and primer details.

<table>
<thead>
<tr>
<th>Gene Fragment</th>
<th>Genome Position&lt;sup&gt;a&lt;/sup&gt;</th>
<th>Primer Name</th>
<th>Primer Sequence (5’→3’)</th>
<th>Amplicon Size (bp)&lt;sup&gt;b&lt;/sup&gt;</th>
<th>Sequence Start 5’</th>
<th>Sequence Start 3’</th>
<th>Sequenced Fragment (bp)&lt;sup&gt;c&lt;/sup&gt;</th>
</tr>
</thead>
<tbody>
<tr>
<td>12S rRNA</td>
<td>639-901</td>
<td>12S Fwd</td>
<td>GTTTATTAAATGCCTTGTCTAAGAA (26)</td>
<td>299</td>
<td>GTCTAAGA</td>
<td>TACGTATT</td>
<td>262</td>
</tr>
<tr>
<td></td>
<td></td>
<td>12S Rvs</td>
<td>GCCCCCAATCAAACATACAA (19)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>9S rRNA</td>
<td>1077-1309</td>
<td>9S Fwd</td>
<td>TGCAATCCGTTAGTTGGGTGA (21)</td>
<td>302</td>
<td>TAAAATCG</td>
<td>TATTATTA</td>
<td>232</td>
</tr>
<tr>
<td></td>
<td></td>
<td>9S Rvs</td>
<td>TCCACACCATTAAATAGCAGCT (22)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CYT b</td>
<td>4126-4733</td>
<td>Sp18 Fwd</td>
<td>GACAGGATTAGAGAGCAGAGAG (23)</td>
<td>717</td>
<td>TTTGTYTT</td>
<td>TAATAYCA</td>
<td>607</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Sp18 Rvs</td>
<td>CAAACCTATCCAAAAAGGATCTG (24)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MURF1</td>
<td>6011-6393</td>
<td>MurfA Fwd</td>
<td>AAGGCRATGGGRATAGWRCCTATAAC (25)</td>
<td>482</td>
<td>ACTAAGYA</td>
<td>ACTTGYTA</td>
<td>382</td>
</tr>
<tr>
<td></td>
<td>6528-6900</td>
<td>MurfA Rvs</td>
<td>TGGAAACATTTATATCATGATTRGGA (26)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>MurfB Fwd</td>
<td>ACMCCCATCCATTCTTCR (18)</td>
<td>423</td>
<td>CAAAAATT</td>
<td>GGATTAT</td>
<td>372</td>
</tr>
<tr>
<td></td>
<td></td>
<td>MurfB Rvs</td>
<td>CTTTGTATATTGTTATGATACRKT (25)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>NDI</td>
<td>7643-8011</td>
<td>NDI Fwd</td>
<td>GCACTTTCTGAAATAATCGAAAA (23)</td>
<td>400</td>
<td>TCGAAAAA</td>
<td>TTGTTAGC</td>
<td>368</td>
</tr>
<tr>
<td></td>
<td></td>
<td>NDI Rvs</td>
<td>TTAACCTTATCGATTGTTTAGCC (25)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>COII</td>
<td>8194-8610</td>
<td>COII Fwd</td>
<td>GTTTATATCTTTTGTGGTTTGTGTG (27)</td>
<td>560</td>
<td>CTTTCTAC</td>
<td>ACCTRCCY</td>
<td>416</td>
</tr>
<tr>
<td></td>
<td></td>
<td>COII Rvs</td>
<td>AAACATGCGATTTATCCATG (22)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ND4</td>
<td>12153-</td>
<td>ND4 Fwd</td>
<td>TTTTGAAGTCTATTTTCCC (23)</td>
<td>302</td>
<td>AATTATAA</td>
<td>CGGTYRTC</td>
<td>239</td>
</tr>
<tr>
<td>Genome position</td>
<td>Amplicon size</td>
<td>Sequence length</td>
<td>Notes</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>-----------------</td>
<td>--------------</td>
<td>----------------</td>
<td>-------</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ND4a Rvs</td>
<td>12392</td>
<td>21</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ND4a Fwd</td>
<td>12153-12392</td>
<td>(23)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ND4a Rvs</td>
<td>TTYTTCCCAATATGTATBGMTAG</td>
<td>502</td>
<td>AATTCTAA</td>
<td>CGGTYRTC</td>
<td>239</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ND4a Rvs</td>
<td>TGTATTAYCGAYCAATTYGCC</td>
<td>(20)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ND5a Fwd</td>
<td>13829-14250</td>
<td>(24)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ND5a Rvs</td>
<td>TATGRTAACAYTTTTGACGAYTCRG</td>
<td>503</td>
<td>GTACATAY</td>
<td>TYTYGTA</td>
<td>421</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ND5a Rvs</td>
<td>GTCCTTCCATGCTYTGC</td>
<td>(19)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ND5b Fwd</td>
<td>14274-14640</td>
<td>(24)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ND5b Rvs</td>
<td>ARAGTACACAGTTGGRTACAYA</td>
<td>444</td>
<td>TGATRCC</td>
<td>GYARACCA</td>
<td>366</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ND5b Rvs</td>
<td>CTTGCAARATACACCCACAAA</td>
<td>(21)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*a* Genome position according to the *T. clypeatus* Sylvio X10/1 reference maxicircle genome (Franzén *et al.*, 2011).

*b* Amplicon size according to *T. clypeatus* Sylvio X10/1. Indels in other strains may cause size variation.

*c* Sequence length according to *T. clypeatus* Sylvio X10/1. Indels in other strains may cause length variation.

*d* An alternate set of *ND4* primers were designed to improve amplification efficiency (Messenger *et al.*, accepted)
2.4.6 Phylogenetic analysis of mtMLST

Sequence data were assembled manually as described for nuclear loci (section 2.4.3). For each isolate, maxicircle sequences were concatenated according to their structural arrangement (12S rRNA, 9S rRNA, CYT b, MURF1, ND1, COII, ND4 and ND5) and in the correct coding direction.

Phylogenetic analysis was performed using ML and Bayesian algorithms, implemented in PhyML (Guindon et al., 2010) or MEGA 5.10 (Tamura et al., 2011) and MrBAYES v3.1 (Ronquist and Huelsenbeck, 2003), respectively. The appropriate nucleotide substitution model for each analysis was selected from 1,624 candidates, based on the Akaike Information Criterion (AIC) or the Bayesian Information Criterion (BIC), respectively, inferred in jMODELTEST 2.1.4 (Darriba et al., 2012). For ML phylogenies, bootstrap support for clade topologies was routinely estimated following the generation of 1000 pseudo-replicated datasets. Bayesian posterior probabilities were derived from multiple independent runs with periodic sampling over a minimum of one million iterations. When molecular dating was undertaken, a Bayesian MCMC method was executed in the Bayesian evolutionary analysis by sampling trees (BEAST) software package (Drummond and Rambaut, 2007).

In incidences when nuclear and mitochondrial phylogenetic incongruence was suspected, alternate tree topologies derived from each dataset were evaluated using Kishino-Hasegawa (KH) (Kishino and Hasegawa, 1989) (to compare two trees) or Shimodaira-Hasegawa likelihood tests (SH) (Shimodaira and Hasegawa, 1999) (to compare >2 trees) implemented in PAML v.4 (Yang, 2007). In selected analyses, haplotype diversity (Hd) was calculated using DnaSP v5.10.1 (Librado and Rozas, 2009).

2.4.7 Multilocus microsatellite typing (MLMT)

Twenty-eight microsatellite loci were routinely used for genotyping, as previously described by Llewellyn et al., 2009a. These markers were distributed across eleven putative chromosomes, including six groups of physically linked loci (Weatherly et al., 2009). Marker choice in each study was determined by amplification reliability and intra-lineage resolution depending on DTU of strains under study.

A full list of microsatellite targets and primers are given in Table 2.5. Five fluorescent dyes were used to label the forward primers: 6-FAM and TET (Proligo, Germany) and NED, PET and VIC (Applied Biosystems, UK) (Figure 2.4).
<table>
<thead>
<tr>
<th>Chromosome</th>
<th>Primer code</th>
<th>Repeat type</th>
<th>Forward/Reverse Primer (5’-3’)</th>
</tr>
</thead>
<tbody>
<tr>
<td>6</td>
<td>6529(CA)$_n$</td>
<td>(CA)$_n$</td>
<td>TGTGAAATGATTTGACCCGA</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>AGAGTCACGCCGCAAAGTAT</td>
</tr>
<tr>
<td>6</td>
<td>6529(TA)$_n$</td>
<td>(TA)$_n$</td>
<td>TGAAGGAGATTTCTCTGCAGG</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>CTCTCATCTTTTGTGTGTCCG</td>
</tr>
<tr>
<td>6</td>
<td>mclF10</td>
<td>(CA)$_n$A(CA)$_n$</td>
<td>GCGTAGCGATATCTTTTC</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>ATCCGCTACCCTATCCAC</td>
</tr>
<tr>
<td>10</td>
<td>6855(TA)(GA)</td>
<td>(TA)$_n$(GA)$_n$</td>
<td>TGATCGACGCACGACATAAT</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>TCCCCATTGCTCTGTTTTAGA</td>
</tr>
<tr>
<td>15</td>
<td>11863(CA)</td>
<td>(CA)$_n$</td>
<td>AGTGGACATCCCCAAAGG</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>CCCTGATGCTGCAGACTCTT</td>
</tr>
<tr>
<td>19</td>
<td>TeUn3</td>
<td>Unknown</td>
<td>CTTAAAGAGATACAAAGAGGAGG</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>CTGTATTTTCAATAACAGG</td>
</tr>
<tr>
<td>19</td>
<td>10101(TA)</td>
<td>(TA)$_n$</td>
<td>AACCCGCGCAGATACATTAG</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>TCCATTGCAAGCAACA</td>
</tr>
<tr>
<td>24</td>
<td>8741(TA)</td>
<td>(TA)$_n$</td>
<td>TGTAACGGAAGGTCTCAGAAT</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>TTGCAGCTTGTTGTCTCCG</td>
</tr>
<tr>
<td>27</td>
<td>10101(TC)</td>
<td>(TC)$_n$</td>
<td>CGTACGACGTTGACAAAC</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>ACAAGTGGTGAGCCAAAG</td>
</tr>
<tr>
<td>27</td>
<td>10101(CA)$_c$</td>
<td>(CA)$_n$</td>
<td>GTGTCGTTGCTCCCCAAAATC</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>AAACTTGCCAAATGTCAGG</td>
</tr>
<tr>
<td>27</td>
<td>10101(CA)$_a$</td>
<td>(CA)$_n$</td>
<td>GTGTCGCTCATGATCAAAC</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>CTGTGGCGAATGTCATAA</td>
</tr>
<tr>
<td>34</td>
<td>6559(TC)</td>
<td>(TC)$_n$</td>
<td>CGCTCCTCAAAGGCACCTTAC</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>ATATGGACGCGTAGGGAGTCG</td>
</tr>
<tr>
<td>37</td>
<td>10187(TTA)</td>
<td>(TTA)$_n$</td>
<td>GAGAGAGATTCGGAACCTAATAG</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>CATGTCCCTTCCCTGGTAAA</td>
</tr>
<tr>
<td>37</td>
<td>10187(CA)(TA)</td>
<td>(CA)$_n$(TA)$_n$</td>
<td>CATGTCATTAAGTGCGGACCAG</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>GCACATGTTGTTGTTGGA</td>
</tr>
<tr>
<td>Chromosomal alignment based on Weatherly et al., 2009.</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>37</td>
<td>10187(TA)</td>
<td>(TA)&lt;sub&gt;n&lt;/sub&gt;</td>
<td>AGAAAAAGTTTACAAACGAGCG</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>CGATGGAGAACGTGAAACAA</td>
</tr>
<tr>
<td>37</td>
<td>10187(GA)</td>
<td>(GA)&lt;sub&gt;n&lt;/sub&gt;</td>
<td>GTCACACCCTAGCGATGACA</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>ACTGCAAAATACCCCTTTG</td>
</tr>
<tr>
<td>37</td>
<td>TcUn2</td>
<td>Unknown</td>
<td>AAACAAATCTAGCGTCTACCATCC</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>GGTGTTGGCGTGTATGATTTG</td>
</tr>
<tr>
<td>37</td>
<td>TcUn4</td>
<td>Unknown</td>
<td>ATGCTCCGAACATATTACTCA</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>GTCGAGCTTCTGTTTGCCT</td>
</tr>
<tr>
<td>39</td>
<td>6925(TG)&lt;sub&gt;n&lt;/sub&gt;</td>
<td>(TG)&lt;sub&gt;n&lt;/sub&gt;</td>
<td>GAAACGCACTCAACCCACAC</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>GGTAGCAAGGCAAAACTTTC</td>
</tr>
<tr>
<td>39</td>
<td>7093(TA)&lt;sub&gt;n&lt;/sub&gt;</td>
<td>(TA)&lt;sub&gt;n&lt;/sub&gt;</td>
<td>GGAACACATCAGCAAGGA</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>AGTGACAAAGGGAGCATTTG</td>
</tr>
<tr>
<td>39</td>
<td>7093(TC)</td>
<td>(TC)&lt;sub&gt;n&lt;/sub&gt;</td>
<td>CCAACATTCAACAAGGGAAA</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>GCATGAATATTGCCGATCT</td>
</tr>
<tr>
<td>39</td>
<td>6925(CT)</td>
<td>(CT)&lt;sub&gt;n&lt;/sub&gt;</td>
<td>CATCAAGAAAAACGAGGA</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>CGGTACCACCTCAAGGAAG</td>
</tr>
<tr>
<td>39</td>
<td>7093(TA)&lt;sub&gt;n&lt;/sub&gt;</td>
<td>(TA)&lt;sub&gt;n&lt;/sub&gt;</td>
<td>CGTGTCCACAGGAGAGA</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>CGTGGAGGAGGATTGA</td>
</tr>
<tr>
<td>39</td>
<td>6925(TG)&lt;sub&gt;n&lt;/sub&gt;</td>
<td>(TG)&lt;sub&gt;n&lt;/sub&gt;</td>
<td>TCGTTCTCTTTACGCTTGCA</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>TAGCAGCAAACAAAAACG</td>
</tr>
<tr>
<td>39</td>
<td>7093(TCC)</td>
<td>(TCC)&lt;sub&gt;n&lt;/sub&gt;</td>
<td>AGACGTTCATATTCGACGCC</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>AGCCACATCCACATTTCCT</td>
</tr>
<tr>
<td>40</td>
<td>11283(TCG)</td>
<td>(TCG)&lt;sub&gt;n&lt;/sub&gt;</td>
<td>ACCACCAGGAGGACGTGAG</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>TGTACAGGGACAGCGAAG</td>
</tr>
<tr>
<td>40</td>
<td>11283(TA)&lt;sub&gt;n&lt;/sub&gt;</td>
<td>(TA)&lt;sub&gt;n&lt;/sub&gt;</td>
<td>AACATCCCACCCACAGGG</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>TTTGAATGCGAGGTTGACA</td>
</tr>
<tr>
<td>41</td>
<td>10359 (CA)(GA)</td>
<td>(CA)&lt;sub&gt;n&lt;/sub&gt;(GA)&lt;sub&gt;n&lt;/sub&gt;</td>
<td>AGTCTACTGCTCCTTGGCA</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>CTGTGGCGAATGCTCATAA</td>
</tr>
</tbody>
</table>
Figure 2.4. Microsatellite primer positions in a 96-well plate.

Amplification reactions were achieved in a final volume of 10µl containing: 1X ThermoPol® Reaction Buffer (New England Biolabs, UK), 4mM MgCl2 (Bioline, UK), 34µM dNTPs (New England Biolabs, UK), 0.75pmol of each primer, 1U Taq DNA polymerase (New England Biolabs, UK) and 1ng of genomic DNA.

PCR reaction conditions for all loci were an initial denaturation step of 4 minutes at 95°C, followed by 30 amplification cycles (95°C for 20 seconds, 57°C for 20 seconds, 72°C for 20 seconds) and a final elongation step at 72°C for 20 minutes. Following amplification, microsatellite products were multiplexed according to Figure 2.5. Allele sizes were determined using a 16-capillary 3730 DNA Analyzer (Applied Biosystems, UK), in conjunction with a fluorescently-tagged size standard (GeneScanTM – 500 LIZ®, Applied Biosystems, UK) and were manually checked for errors in GeneMapper® software v3.7. All isolates were typed ‘blind’ to control for subjective user bias (Figure 2.6). Additional meta-data from previous microsatellite publications were included in selected analyses.
Figure 2.5. Schematic of multiplexing microsatellite PCR products.

Figure 2.6. Example of an ambiguous microsatellite profile for strains from different *T. cruzi* DTUs (A: Arma13 cl1; TcIII and B: Bug2148 cl1; TcV). In A the second peak was ambiguous according to the software and manually edited by the user.
2.4.8 Phylogenetic analysis of MLMT

Individual-level sample clustering was routinely defined using a NJ tree based on pair-wise distances ($D_{AS}$: 1 – proportion of shared alleles at all loci/n) between microsatellite genotypes calculated in MICROSAT v1.5d (Minch et al., 1997) under the infinite-alleles model (IAM). To accommodate multi-allelic genotypes (≥3 alleles per locus), a script (kindly provided by M. Llewellyn) was written in Microsoft Visual Basic to generate random multiple diploid re-samplings of each multilocus profile. For each microsatellite dataset, a NJ phylogenetic tree was constructed in PHYLIP v3.67 (Felsenstein, 1989) using a pair-wise distance matrix derived from the mean across multiple re-sampled datasets. Majority rule consensus analysis of 10,000 bootstrap trees was performed in PHYLIP v3.67 by combining 100 bootstraps generated in MICROSAT v1.5d, each drawn from 100 respective randomly re-sampled datasets.

In selected studies, putative subpopulations were defined using a nonparametric approach (free from Hardy-Weinberg assumptions). A K-means clustering algorithm, implemented in adegenet (Jombart et al., 2008), was used to determine the optimal number of ‘true’ populations, by reference to the BIC, which reaches a minimum when approaching the best supported assignment of individuals to the appropriate number of clusters. The relationship between these clusters and the individuals within them was then evaluated via a discriminant analysis of principal components (DAPC) according to Jombart et al., 2010.

Population-level genetic diversity was evaluated using sample size corrected allelic richness ($A_r$) in FSTAT 2.9.3.2 (Goudet, 1995). Mean $F_{IS}$, which measures the distribution of heterozygosity within and between individuals, was calculated per population in FSTAT 2.9.3.2. $F_{IS}$ can vary between -1 (all loci heterozygous for the same alleles) and +1 (all loci are homozygous for different alleles). $F_{IS} = 0$ indicates Hardy-Weinberg allele frequencies. Sample size corrected private (population-specific) allele frequency per locus (PA/L) was calculated in HP-Rare (Kalinowski, 2005).

Population subdivision was estimated using pair-wise $F_{ST}$, linearised with Slatkin’s correction, in ARLEQUIN v3.11 (Excoffier et al., 2005). Statistical significance was assessed via 10,000 random permutations of alleles between populations. Within population subdivision was evaluated in ARLEQUIN v3.11 using a hierarchal Analysis of Molecular Variance (AMOVA). Population-level heterozygosity indices were calculated in ARLEQUIN v3.11 and associated significance levels for $p$-values derived after performing a sequential Bonferroni correction to minimise the likelihood of Type 1 errors (Rice, 1989).

Multilocus linkage disequilibrium, estimated by the Index of Association ($I_A$) was calculated in MULTILOCUS 1.3b (Agapow and Burt, 2001) and statistical significance evaluated by comparison to a null distribution of 1000 randomisations. Mantel’s tests for the effect of isolation by distance within populations (pair-wise genetic vs. geographic distance) were implemented in GenAIEx 6.5 using 10,000 random permutations (Peakall and Smouse, 2012), as indicated.
2.4.9 Illumina amplicon sequencing
Two 450 bp multi-copy polymorphic sequence markers (TcGP63 and ND5) were chosen to investigate intra-host parasite multiclonality. Degenerate primers were designed for TcGP63 family surface proteases (Cuevas et al., 2003) by reference to sequences retrieved from TriTypDB for Sylvio X10/1 (TcI), JR cl4 (TcI), Esm cl3 (TcII) and CL Brener (TcVI) (Aslett et al., 2010). TcGP63 sequences were aligned in MUSCLE (Edgar, 2004) and primers were manually designed to target an invariable flanking region (TcGP63_F CCAGYTGGTGTAATRCTGCYGCC and TcGP63_R RGAACCGATGTCATGGGGCAA). Degenerate primers to amplify ND5 (fragment b) were previously designed by Messenger et al., 2012 (Table 2.4).

PCR reactions were performed using the FastStart High Fidelity PCR System (Roche, UK), according to the manufacturer’s instructions. Amplifications were undertaken using the Fluidigm® platform with a reduction of the manufacturer’s recommended number of cycles to 26 to minimise PCR amplification bias.

PCR reactions were performed in triplicate for each patient sample and pooled, prior to sequencing. Negative controls were included in all PCR and sequencing steps. Equimolar concentrations of ND5 and TcGP63 amplicons from 96 DNA samples were multiplexed on Illumina runs using dual index sequence tags (Illumina Inc). Sequencing was undertaken using a MiSeq platform with 2 x 250 bp read lengths (Reagent Kit version 2), according to the manufacturer’s protocol.

In addition to clinical isolates, a dilution series of control samples were also sequenced. Controls comprised artificial mixes of DTUs I-VI genomic DNA at equimolar concentrations.

2.4.10 Illumina amplicon sequence data analysis
De-multiplexed paired-end sequences were submitted to quality control and trimming in Sickle (Joshi and Fass, 2011) and mate pairs were trimmed in FASTX Toolkit (http://hannonlab.cshl.edu/fastx_toolkit/). ND5, TcGP63 and contaminating sequences were then sorted against a reference using Bowtie 2 (Langmead and Salzberg, 2012). Individual paired reads were found to be overlapping in only a minority of cases; a truncated central fragment was used in all downstream analyses.

Further sequence manipulations were undertaken using FASTX Toolkit and custom awk scripts (kindly provided by M. Llewellyn) to parse files and concatenate mate pairs for downstream analysis. MUSCLE (Edgar, 2004) was used for alignment of sequences in each sample. Next, analysis was undertaken in the Mothur software package (Schloss et al., 2009) to eliminate putative PCR chimeras and examine clustering of individual sequences.

The Shannon index of diversity was calculated at the intra-patient level based on STs defined at 97% and 99% identity cut-offs in Mothur. Comparisons among patient cohorts were made
via analyses of covariance and linear regression in R (http://CRAN.R-project.org). Sequence datasets for patients from each cohort were then merged and analyses conducted using 97% (ND5) and 99% (TcGP63) sequence clusters defined with UPARSE (Edgar, 2013). Weighted UniFrac distances between STs among samples were generated and subsequently clustered via a principal coordinates analysis in QIIME (Caporaso et al., 2010). Estimates of diversifying selection among TcGP63 STs were made in KaKs Calculator (Zhang et al., 2006) using Yang and Neilson’s 2000 approximate method (Yang and Nielsen, 2000) and tested for significance using a Fisher’s exact test. To test across putative TcGP63 gene families (97% cut-off), 99% STs within each cohort population were pooled. Secondly, tests were conducted among STs within each 97% category per cohort.
3. Development and evaluation of *T. cruzi* phylogenetic markers for DTU-level genotyping and higher resolution population genetics

3.1 Background

3.1.1 A brief history of *T. cruzi* taxonomy

Elucidating the complex epidemiology, clinical variability and phylogeography underlying Chagas disease requires a clear understanding of the parasite’s genetic diversity (Campbell *et al.*, 2004; Miles *et al.*, 2009). Historically, the taxonomy of *T. cruzi* has been hindered by a lack of standardized molecular typing methods and the use of various alternative nomenclatures (recently reviewed by Zingales *et al.*, 2012). The earliest attempts to characterize *T. cruzi* strain variation, based on multilocus enzyme electrophoresis (MLEE), classified isolates into three major groups or ‘zymodemes I, II and III’ (Miles *et al.*, 1977; 1978), which were later subdivided into 43 ‘clonets’ (Tibayrenc and Ayala, 1988) (Table 3.1).

Subsequent genotyping of additional strains using MLEE (Tibayrenc *et al.*, 1993), random amplification of polymorphic DNA (RAPD) (Steindel *et al.*, 1993) and nuclear loci (Souto *et al.*, 1996; Fernandes *et al.*, 1998a, 1998b), grouped isolates into two major lineages, designated *T. cruzi* I and *T. cruzi* II (Anonymous, 1999) (Table 3.1). More recently, using MLST, TcII was divided into TcIIa-e (Brisse *et al.*, 2000, 2001), which were latterly renamed to TcII-TcVI by international consensus to remove any presumptive sub-lineage designations (Zingales *et al.*, 2009).

Each of the six *T. cruzi* lineages is considered a discrete typing unit (DTU), defined as “a collection of strains that are genetically more closely related to each other than to any other strain and that share one or several specific characters” (Tibayrenc, 1998). However, the criteria for division, number of subgroups, and their precise biological and evolutionary relevance are still a popular subject for debate (Devera *et al.*, 2003; Herrera *et al.*, 2007b; Guhl and Ramírez, 2011). For example, there has been significant interest in classifying Tcl isolates on the basis of either transmission cycle (Tcl-a-e) (Herrera *et al.*, 2007b, 2009; Falla *et al.*, 2009; Cura *et al.*, 2010) or association with human infection (TclDOM) (Zumaya-Estrada *et al.*, 2012; Ramírez *et al.*, 2012). Furthermore, a potential seventh DTU (TcBat) has recently been identified from *Chiroptera* species in northern South and Central America (Marcili *et al.*, 2009; Pinto *et al.*, 2012; Ramírez *et al.*, 2014).
Table 3.1. Comparison of *T. cruzi* historical and contemporary nomenclatures.

<table>
<thead>
<tr>
<th>Term</th>
<th>Technique</th>
<th>I</th>
<th>II</th>
<th>III</th>
<th>IV</th>
<th>V</th>
<th>VI</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>DTU (Current Nomenclature)</td>
<td>DTU: 24Sa rRNA; SSU rDNA; Cytb; Histone I2B; ITS1 rDNA; 18S rRNA; gGAPDH</td>
<td>I</td>
<td>II</td>
<td>III</td>
<td>IV</td>
<td>V</td>
<td>VI</td>
<td>Zingales <em>et al.</em>, 2009</td>
</tr>
<tr>
<td></td>
<td>SL-IR</td>
<td>Ia, Ib, Ic, Id, Ie</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>TcBat</td>
</tr>
<tr>
<td>Putative Intra-TcI Subdivisions</td>
<td>MLMT; mtMLST</td>
<td>I_{DOM}</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Herrera <em>et al.</em>, 2007b; 2009; Falla <em>et al.</em>, 2009; Cura <em>et al.</em>, 2010</td>
</tr>
<tr>
<td>DTU'</td>
<td>MLEE; RAPD; 24Sa rRNA: 18S rRNA</td>
<td>I</td>
<td>IIb</td>
<td>IIc</td>
<td>IIa</td>
<td>IId</td>
<td>Ile</td>
<td>Brisse <em>et al.</em>, 2000, 2001</td>
</tr>
<tr>
<td></td>
<td>MLEE</td>
<td>ZB</td>
<td>ZA</td>
<td>ZC</td>
<td>ZD</td>
<td>ZB</td>
<td>ZB</td>
<td>Romanha, 1982; Carneiro <em>et al.</em>, 1990</td>
</tr>
<tr>
<td></td>
<td>ITS-RFLP; 24Sa rRNA</td>
<td>ZIII-A</td>
<td></td>
<td>ZIII-B</td>
<td></td>
<td></td>
<td></td>
<td>Mendonça <em>et al.</em>, 2002</td>
</tr>
<tr>
<td>Biodemes</td>
<td>MLEE</td>
<td>III</td>
<td>II</td>
<td></td>
<td>I</td>
<td>I</td>
<td></td>
<td>Andrade and Magalhães, 1997</td>
</tr>
<tr>
<td>Clonet</td>
<td>MLEE</td>
<td>1-25</td>
<td>30-34</td>
<td>35-37</td>
<td>26-29</td>
<td>38, 39</td>
<td>40-43</td>
<td>Tibayrenc and Ayala, 1988</td>
</tr>
<tr>
<td>------------------</td>
<td>-----------</td>
<td>-------------</td>
<td>----------</td>
<td>----------</td>
<td>----------</td>
<td>----------------</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Lineage</strong></td>
<td>24Sα rRNA; SL-IR; RAPD</td>
<td>2</td>
<td>1</td>
<td>1/2</td>
<td>1/2</td>
<td>Souto <em>et al.</em>, 1996</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>24Sα rRNA; SL-IR</td>
<td>2</td>
<td>1</td>
<td>2’</td>
<td>2’</td>
<td>Fernandes <em>et al.</em>, 1998a; Fernandes <em>et al.</em>, 1998b</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Clade</strong></td>
<td>TR; DHFR-TS; COII-ND1</td>
<td>A</td>
<td>C</td>
<td>B</td>
<td>D</td>
<td>B+C</td>
<td>Machado and Ayala, 2001</td>
<td></td>
</tr>
<tr>
<td><strong>Riboclades</strong></td>
<td>24Sα rRNA; 18S rRNA</td>
<td>2</td>
<td>1</td>
<td>3</td>
<td>4</td>
<td>3</td>
<td>de Freitas <em>et al.</em>, 2006</td>
<td></td>
</tr>
<tr>
<td><strong>Haplogroups</strong></td>
<td>MLMT; 24Sα rRNA; COII; ND1; Cytb</td>
<td>ZZ</td>
<td>YY</td>
<td>XX</td>
<td>XY</td>
<td>XY</td>
<td>Augusto-Pinto <em>et al.</em>, 2003</td>
<td></td>
</tr>
<tr>
<td></td>
<td>TcMSH2</td>
<td>A</td>
<td>C</td>
<td></td>
<td>B</td>
<td></td>
<td>Pedroso <em>et al.</em>, 2007</td>
<td></td>
</tr>
<tr>
<td><strong>Group</strong></td>
<td>Karyotyping</td>
<td>C</td>
<td>B</td>
<td>D</td>
<td>A</td>
<td>A</td>
<td>Nunes <em>et al.</em>, 1997</td>
<td></td>
</tr>
<tr>
<td><strong>SL-IR</strong></td>
<td>II</td>
<td>1</td>
<td>I</td>
<td>I</td>
<td>I</td>
<td>I</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Reference Strain</strong></td>
<td>Sylvio X10/1</td>
<td>Esm cl3</td>
<td>M5631 cl5</td>
<td>CanIII cl1</td>
<td>Sc43 cl1</td>
<td>CL Brener</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

3.1.2 Contemporary T. cruzi genotyping: perils and pitfalls

Establishing any relationship between T. cruzi genotype and clinical outcome, ecological niche, host association, geographical distribution etc., is complicated by a number of inherent biological features relating to both parasite infection dynamics and our current repertoire of genotyping techniques. In humans (Vago et al., 2000; Burgos et al., 2008; 2010; Ramírez et al., 2012), triatomine bugs (Bosseno et al., 1996; Yeo et al., 2007; Cardinal et al., 2008) and mammalian reservoir hosts (Yeo et al., 2005; Llewellyn et al., 2011; Rocha et al., 2013), mixed infections of distinct parasite clones are not exceptional but inevitable. Among highly endemic areas, long-term inhabitants are repeatedly infected by multiple contacts with different triatomines (Nouvellet et al., 2013), which in turn have fed on various infected humans and/or mammals, depending on the local disease ecology.

Levels of intra-host parasite multiclonality might be expected to increase proportionally to vector exposure. However, this assumes a constant force of infection, incomplete cross-genotypic immunity, and lack of genotype interaction (e.g. genotype displacement, reciprocal inhibition, potentiation or recombination (Pinto et al., 1998; da Silveira Pinto et al., 2000; Gaunt et al., 2003; Martins et al., 2007; Araújo et al., 2014), transmission population bottlenecks (as observed in related trypanosomes (Oberle et al., 2009)) or any additional mechanisms that prevent the establishment of secondary infections. The complexity of natural multiclonal parasite populations is largely unknown and the ability to detect them restricted by genetic marker resolution (Llewellyn et al., 2011; Valadares et al., 2012). The study of this phenomenon conventionally necessitates deriving biological clones from live parasite populations, prior to genetic typing, which introduces a range of potential adaptation biases, discussed below.

Genotyping of natural T. cruzi strains can be performed either directly from field samples (blood, tissue biopsies or vector faeces) or following parasite isolation by hemoculturing or xenodiagnoses. Due to the scarcity of parasites in peripheral blood, especially in chronically infected patients, the former method has limited sensitivity. The principal drawback associated with parasite isolation is selection bias for particular subpopulations, initially by preferential outgrowth due to faster dividing rates and/or culture media (Dvorak et al., 1980; Alves et al., 1994; Devera et al., 2003) and subsequently by loss of clonal diversity from serial maintenance in axenic culture or animals (Engel et al., 1982; Deane et al., 1984b; 1984c; 1984d; Morel et al., 1986; Alves et al., 1993). Hemoculturing is laborious and recovery rates (usually less than 30% among chronic patients (Siriano et al., 2011)), entirely determined by size of parasite inoculum and distribution within the starting sample. Xenodiagnosis, which can facilitate greater parasite recovery, has also been shown to vary depending on vector permissibility to local strains (Miles et al., 1984; Luquetti et al., 1986). Furthermore, due to differential tropisms of strains, circulating clones which are isolated by hemoculture/xenodiagnosis are often genetically distinct from those sequestered in tissues (Vago et al., 2000; Burgos et al., 2008; 2010) and even between sequential blood samples (Sánchez et al., 2013). Together these observations strongly suggest that intra-host parasite diversity is routinely underestimated.
A plethora of molecular genotyping techniques have been developed to characterize *T. cruzi* genetic diversity, with varying degrees of resolution, experimental ease, reproducibility, subjectivity and transferability (Table 3.2). Typing of genetic polymorphisms in conserved housekeeping genes can define major genetic lineages (Souto et al., 1996; Fernandes et al., 1998a; 1998b; Brisse et al., 2000; 2001), while analysis of hypervariable loci such as microsatellites (Oliveira et al., 1998; 1999; Llewellyn et al., 2009a; 2009b), or kDNA minicircles (Morel et al., 1980; Veas et al., 1990; Telleria et al., 2006; Velázquez et al., 2008), potentially allows identification of profiles specific to individual strains. Choice of typing methodology is principally determined by sample source, research question and laboratory resources.

Direct clinical genotyping is currently based on size polymorphisms in multi-copy genetic markers, including the nuclear spliced-leader intergenic region (*SL-IR*), 24α rDNA (Souto et al., 1996) and 18SrDNA (Brisse et al., 2001), and minicircle sequences (Lages-Silva et al., 2006; Burgos et al., 2007; 2008; 2010) (for more detailed descriptions of historical genotyping techniques see Macedo et al., 2002; 2004; Devera et al., 2003). One major confounder associated with the use of any multi-copy gene is the level of intra-clone copy number and position homology to ensure comparability between strains; genome size (Dvorak et al., 1982; Lewis et al., 2009b), karyotype (Henriksson et al., 1990; 1993; Vargas et al., 2004; Souza et al., 2011; Lima et al., 2013) and chromosomal arrangements of tandem repeat regions (Wagner and So, 1990; O’Connor et al., 2007), are known to differ widely between natural *T. cruzi* strains and even biological clones derived from the same population. Similar caveats affect minicircle-based genotyping, which vary in copy number and complement between major DTUs (Morel et al., 1980; Moreira et al., 2013), are susceptible to contamination (Schijman et al., 2011) and whose profiles are highly sensitive to minor changes in reaction conditions, raising issues of reproducibility (Alves et al., 1993; Segatto et al., 2013). With all of these methods, strain DTU assignment is often dependent on absence of PCR products/restriction fragment bands which cannot be distinguished from variation in as yet untested strains.

One recent method, developed to circumvent some of the limitations associated with clinical genotyping is to adopt an indirect approach, exploiting serological detection of antibodies produced in response to DTU-specific *T. cruzi* antigens (Di Noia et al., 2002; Mendes et al., 2013; Bhattacharyya et al., 2014). However, serology cannot distinguish between contemporary infection or historical exposure to a particular lineage, given anti-*T. cruzi* antibodies can take years to decline.
Table 3.2. Overview of current and historical *T. cruzi* genotyping methods.

<table>
<thead>
<tr>
<th>Genotyping Method</th>
<th>Method Description</th>
<th>Example of Genetic Markers</th>
<th>Reproducibility b/w Assays</th>
<th>Level of Resolution</th>
<th>Reagent Cost</th>
<th>Advantages</th>
<th>Disadvantages</th>
<th>References</th>
</tr>
</thead>
<tbody>
<tr>
<td>MLEE</td>
<td>Measures differences in electrophoretic mobilities of isoenzymes</td>
<td>ASAT, ALAT, PGM, ACON, MPI, ADH, MDH, ME, ICD, 6PGD, G6PD, GD, PEP, GPI</td>
<td>High</td>
<td>DTU-level</td>
<td>Moderate</td>
<td>- Easy visual interpretation&lt;br&gt;- Data amenable to numerical taxonomic analysis, e.g. rates of similarity or genetic distance</td>
<td>- Requires large quantities of parasite lysate from live strains</td>
<td>Ready and Miles, 1980; Barratt <em>et al.</em>, 1980; Miles <em>et al.</em>, 1981a; 1984; Romanha, 1982; Carneiro <em>et al.</em>, 1990</td>
</tr>
<tr>
<td>RAPD</td>
<td>Short random sequence primers used to amplify unknown DNA fragments to create unique band patterns</td>
<td>N/A</td>
<td>Low</td>
<td>DTU-level</td>
<td>Low</td>
<td>- Can be performed directly on field samples&lt;br&gt;- No prior sequence knowledge needed&lt;br&gt;- Unlimited number of primers&lt;br&gt;- Data amenable to numerical taxonomic analysis</td>
<td>- Reproducibility issues&lt;br&gt;- Dominant markers may conceal heterozygosity&lt;br&gt;- Strain profiles may vary with DNA template amount and quality</td>
<td>Steindel <em>et al.</em>, 1993; Souto <em>et al.</em>, 1996; Brisse <em>et al.</em>, 2000, 2001</td>
</tr>
<tr>
<td>kDNA-RFLP</td>
<td>Restriction fragment length analysis of kinetoplast minicircle hypervariable region (mHVR)</td>
<td>mHVR</td>
<td>Low</td>
<td>Intra-lineage</td>
<td>Low</td>
<td>- Hypervariable markers&lt;br&gt;- Can produce strain-specific profiles</td>
<td>- Requires isolation of kDNA from live parasites&lt;br&gt;- Strain profile inheritance may not be stable or correlate with nuclear typing</td>
<td>Morel <em>et al.</em>, 1980</td>
</tr>
<tr>
<td>Method</td>
<td>Analysis of mHVR by radioactive probe hybridization</td>
<td>mHVR</td>
<td>Low</td>
<td>DTU-level</td>
<td>Intra-lineage</td>
<td>mHVR</td>
<td>Low</td>
<td>DTU-level</td>
</tr>
<tr>
<td>-------------------------</td>
<td>-----------------------------------------------------</td>
<td>------</td>
<td>-----</td>
<td>-----------</td>
<td>--------------</td>
<td>------</td>
<td>-----</td>
<td>-----------</td>
</tr>
<tr>
<td>kDNA hybridization</td>
<td>Comparison of chromosome size variation by PFGE separation and radioactive probe hybridization</td>
<td>1F8, cruzipan, FFAg6, Tc2, CA7.12, CA7.32, P19</td>
<td>Moderate</td>
<td>DTU-level</td>
<td>Moderate</td>
<td>Moderate</td>
<td>DTU-level</td>
<td>Moderate</td>
</tr>
<tr>
<td>Karyotyping (aCSDI)</td>
<td>Analysis of variability in nuclear minisatellites by restriction digestion and probe hybridization</td>
<td>1F8, cruzipan, FFAg6, Tc2, CA7.12, CA7.32, P19</td>
<td>Moderate</td>
<td>DTU-level</td>
<td>Moderate</td>
<td>Moderate</td>
<td>DTU-level</td>
<td>Moderate</td>
</tr>
<tr>
<td>DNA Fingerprinting</td>
<td>Analysis of size polymorphisms in mHVR amplified by low stringency primers</td>
<td>1F8, cruzipan, FFAg6, Tc2, CA7.12, CA7.32, P19</td>
<td>Moderate</td>
<td>DTU-level</td>
<td>Moderate</td>
<td>Moderate</td>
<td>DTU-level</td>
<td>Moderate</td>
</tr>
<tr>
<td>LSSP-PCR</td>
<td>Analysis of size polymorphisms in mHVR amplified by low stringency primers</td>
<td>mHVR</td>
<td>Low</td>
<td>DTU-level</td>
<td>Low</td>
<td>- Highly sensitive</td>
<td>Low</td>
<td>- Hypervariable markers</td>
</tr>
</tbody>
</table>

- Potential contamination problems due to very high copy number
- DNA probes may cross-react b/w DTUs
- Can produce strain-specific profiles
- Potential contamination problems due to very high copy number
<table>
<thead>
<tr>
<th>Method</th>
<th>Analysis of size polymorphisms in multicopy gene fragments</th>
<th>PCR Product Size Polymorphism</th>
<th>PCR-RFLP</th>
<th>Nucleotide sequencing: nuclear loci (nMLST)</th>
<th>Nucleotide sequencing: mitochondrial gene</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>SL-IR, 24Sa rRNA, 18S rRNA</td>
<td>High</td>
<td>High</td>
<td>High (Intra-lineage)</td>
<td>High (DTU-level)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>DTU-level</td>
<td>Moderate</td>
<td>High</td>
<td>High</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Low</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>DTU assignment based on presence/absence of amplicons; insensitive to mutations in untested strains</td>
<td></td>
<td>DTU assignment based on presence/absence of SNPs; insensitive to mutations in untested strains</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Can be performed directly on field samples</td>
<td></td>
<td>Can be performed directly on field samples</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Requires limited technical expertise</td>
<td></td>
<td>Requires limited technical expertise</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>DTU assignment based on presence/absence of amplicons; insensitive to mutations in untested strains</td>
<td></td>
<td>DTU assignment based on presence/absence of amplicons; insensitive to mutations in untested strains</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Unknown intra-strain copy homology</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Souto et al., 1996; Fernandes et al., 1998a; Fernandes et al., 1998b</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Rozas et al., 2007; Lewis et al., 2009a; Van der Auwera et al., submitted; Cosentino and Agüero, 2012</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Machado and Ayala, 2001; Augusto-Pinto et al., 2003; Yeo et al., 2011; Lauthier et al., 2012</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>de Freitas et al., 2006; Carranza</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

- kDNA signatures may vary with DNA template amount and quality
- Can be performed directly on field samples
- Requires limited technical expertise
- DTU assignment based on presence/absence of amplicons; insensitive to mutations in untested strains
- Unknown intra-strain copy homology
- Requires live strains
- Level of intra-lineage resolution dependent upon analysis of multiple loci
- Requires live strains
- Data amenable to MLST analysis
- Data highly reproducible, portable and transferable b/w laboratories
- Requires live strains
- Data amenable to MLST analysis
- Requires live strains
- Data amenable to MLST analysis
- Requires live strains
- Data amenable to MLST analysis
- Requires live strains
<table>
<thead>
<tr>
<th>Maxicircle loci (mtMLST)</th>
<th>Fragments</th>
<th>MURF1, ND1, COII, ND4, ND5, ND7</th>
<th>Intra-lineage</th>
<th>MLST analysis</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>- Data highly</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>reproducible,</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>portable and</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>transferable b/w</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>laboratories</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>- Potential</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>phylogenetic</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>incongruence</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>with nuclear</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>loci</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>- Identifies</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>3 maxicircle</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>classes (TcI,</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>TcII and</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>TcIII-VI); not</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>specific to</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>all 6 DTUs</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Nucleotide sequencing:</td>
<td>SNP analysis in</td>
<td>mHVR</td>
<td>High</td>
<td>DTU-level</td>
</tr>
<tr>
<td>minicircle regions</td>
<td>mHVR</td>
<td></td>
<td>High</td>
<td>High</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>- Hypervariable</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>markers</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>- Can produce</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>strain-</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>specific profiles</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>- Strain profile</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>may not</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>be DTU-specific;</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>minor</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>sequence classes</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>shared</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>b/w DTUs</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>FFLB</td>
<td>Analysis of size</td>
<td>28Sα rRNA, 18S</td>
<td>High</td>
<td>DTU-level</td>
</tr>
<tr>
<td></td>
<td>polymorphisms in</td>
<td>rRNA</td>
<td></td>
<td>High</td>
</tr>
<tr>
<td></td>
<td>multicopy gene</td>
<td>fragments</td>
<td></td>
<td>- Can be performed</td>
</tr>
<tr>
<td></td>
<td>fragments</td>
<td></td>
<td></td>
<td>directly on field</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>samples</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>- Unable to differentiate</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>hybrid lineages (TcV</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>and TcVI)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>HMR</td>
<td>Analysis of amplicon melting</td>
<td>SL-IR, 24Sα rRNA</td>
<td>High</td>
<td>DTU-level</td>
</tr>
<tr>
<td></td>
<td>temperatures generated by real-time PCR</td>
<td></td>
<td></td>
<td>Moderate</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>- Data rapidly</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>generated in real time</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>- Requires live</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>strains</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>- Difficult to standardize</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>b/w laboratories</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>- Requires specialized</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>laboratory infrastructure</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MLMT</td>
<td>Analysis of allele size differences in microsatellite repeat regions</td>
<td>10101(CA), 11283(TA), 7093(TA), TeUn4, mclf10, 10187(CA)(TA), 6855(TA)(GA), 10359(CA), 8741(TA),</td>
<td>Moderate</td>
<td>DTU-level</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Intra-lineage</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>High</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>- Neutrally evolving,</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>co-dominant,</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>hypervariable markers</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>- Can produce</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>strain-</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>specific MLGs</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>- Requires live strains</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>- Prone to homoplasy</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>- Data interpretation</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>highly subjective</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
AmpliSeq Sequencing

| Analysis of millions of sequencing reads generated by Illumina deep sequencing | TcGP63, ND5 | High | DTU-level Intra-lineage Parasite multiclonality | Very high | - Can detect intra-host parasite multiclonality and genetic diversity | - Requires live strains - Prone to loss of clonal diversity from parasite isolation - Requires bioinformatics expertise, computational infrastructure and comparatively high cost reagents | Llewellyn et al., accepted |
Additional genotyping options are available for axenic parasite cultures, including karyotyping (Henriksson et al., 2002; Pedroso et al., 2007), DNA fingerprinting (Macedo et al., 1992) and microsatellite analyses (Llewellyn et al., 2009a; 2009b; Ocaña-Mayorga et al., 2010; Ramírez et al., 2012). However, to date, no single, widely validated genetic marker affords complete, unequivocal DTU resolution (Cosentino and Agüero, 2012) and reliance on only one target is inadvisable given the potential confounding influence of genetic exchange (Westenberger et al., 2005). The availability of reference whole genome sequences (El-Sayed et al., 2005; Weatherly et al., 2009; Aslett et al., 2010; Franzen et al., 2011) has re-invigorated interest in exploring T. cruzi genetic diversity and encouraged the development of nucleotide sequence-based genotyping techniques, particular MLST, and more recently Illumina amplicon deep sequencing.

MLST was originally developed to characterize bacterial (Maiden et al., 1998; Spratt, 1999; Dingle et al., 2001; Enright et al., 2001; Nallapareddy et al., 2002) and fungal (Bougnoux et al., 2003; Tavanti et al., 2003; 2005) species and involves sequencing short fragments of usually seven or more chromosomally-independent, single-copy, housekeeping genes. Sequence polymorphisms within each locus are categorized as distinct STs, which can be concatenated to produce a unique allelic profile per strain. MLST data have the advantages of being highly reproducible, minimally subjective, electronically transferable between laboratories (especially if deposited in international databases such as http://pubmlst.org and http://www.mlst.net) and amenable to a range of different population genetics analyses.

With the advent of next generation sequencing technologies and concomitant plummeting reagent costs, deep sequencing approaches are increasingly used to explore intra-host pathogen genetic diversity (McElroy et al., 2014). Illumina amplicon sequencing generates millions of ‘short’ sequencing reads from individual samples, potentially allowing correlation of read depth with genotype abundance within multiclonal populations. This strategy has been used to examine natural multiplicity of infection in Plasmodium falciparum (Juliano et al., 2010; Manske et al., 2012; Taylor et al., 2013), the dynamics of HIV anti-retroviral escape mutations, prior to the emergence of clinical drug resistance (Wang et al., 2007; Gibson et al., 2014), as well as expose serial population bottlenecks in Trypanosoma brucei brucei transmission (Oberle et al., 2009).
3.2 Objectives
The aim of this chapter was to develop and evaluate nuclear and mitochondrial phylogenetic markers for *T. cruzi* DTU-level assignment, higher resolution population genetics studies and the investigation of intra-host parasite multiclonality.

Specific objectives were to:

a. Critically assess the suitability of candidate nuclear housekeeping genes for inclusion in an MLST scheme for *T. cruzi* DTU-level assignment and intra-lineage diversity studies.

b. Contribute to the formalization of a standardized nuclear MLST scheme (nMLST), validated across a cohort of reference strains, representative of all six DTUs.

c. Assemble mitochondrial Illumina sequencing reads from the TcI reference strain Sylvio X10/1 to resolve the existence of mitochondrial heteroplasmy, a potential genotyping confounder.

d. Optimize a panel of mitochondrial multilocus sequence targets (mtMLST) to describe TcI intra-lineage genetic diversity.

e. Compare phylogenetic incongruence between nuclear and mitochondrial topologies to uncover novel mitochondrial introgression events occurring within natural populations of TcI.

f. Develop nuclear and mitochondrial markers to characterize intra-host parasite multiclonality by Illumina amplicon deep sequencing.
3.3 Results

3.3.1 Development and evaluation of nuclear MLST targets

To develop an MLST scheme for *T. cruzi*, nine nuclear housekeeping targets were selected from published literature and with reference to available genome sequences (Weatherly *et al.*, 2009; Aslett *et al.*, 2010). Candidate genes were chosen on the basis of being single-copy, chromosomally-independent and under stabilising selection with a ratio of non-synonymous to synonymous amino acid changes (dN/dS) less than one (Maiden, 2006; Odds and Jacobson, 2008).

The discriminatory power (number of genotypes identified per total number of isolates) and typing efficiency (number of genotypes identified per polymorphic site) of each target was assessed across a panel of 39 cloned reference strains, representing genetic and geographical diversity of all six *T. cruzi* DTUs. Sequence data were then concatenated to produce unique allelic profiles per strain (diploid sequence types; DST) and the minimum combination of genes was derived by subtractive analysis to: (i) assign isolates to DTU-level; and (ii) enable higher resolution intra-lineage analysis.

This study is reported in full below by Yeo *et al.*, 2011.

In summary:

- Robust amplification was confirmed for all targets and all genes satisfied the criteria for MLST candidates, with the exception of two loci (*TcAPX* and *TR*), found to be under positive selection (dN/dS >1).
- A minimum panel of four MLST targets (*DHFR-TS*, *Met-III*, *RB19* and *TcGPXII*) can be used to unequivocally assign isolates to DTU-level. No single gene was able to differentiate all 39 reference strains.
- An expanded panel of six MLST targets (*DHFR-TS*, *Met-II*, *Met-III*, *RB19*, *TcMPX* and *TR*) afforded the greatest discriminatory power (DP=0.97), distinguishing 38 out of 39 reference isolates. This DP was equivalent to that obtained using all nine concatenated loci.
- Three genes (*LYT1*, *DHFR-TS* and *RB19*) separated genetically homogeneous DTUs TcV and TcVI with significant bootstrap support. For a subset of targets (*Met-II*, *Met-III*, *RB19*, *TcMPX*), loss of heterozygosity (LOH) affecting restricted portions of chromosomes, was observed in both TcV and TcVI, cautioning the solitary use of these loci for lineage assignment of hybrid strains.
- Inter- and intra-lineage phylogenetic incongruence between individual genes trees was indicative of historical nuclear recombination.
- MLST provided a reliable and reproducible method to characterize parasite strains, with the potential to contribute substantially to our understanding of *T. cruzi* genetic diversity, following the establishment of a standardized protocol.
Please be aware that one cover sheet must be completed for each ‘Research Paper’ included in a thesis.

1. For a ‘research paper’ already published
   1.1. Where was the work published? PLoS Neglected Tropical Diseases
   1.2. When was the work published? 2011
   1.2.1. If the work was published prior to registration for your research degree, give a brief rationale for its inclusion
   N/A
   1.3. Was the work subject to academic peer review? Yes
   1.4. Have you retained the copyright for the work? Yes/No
      If yes, please attach evidence of retention.
      If no, or if the work is being included in its published format, please attach evidence of permission from copyright holder (publisher or other author) to include work

2. For a ‘research paper’ prepared for publication but not yet published
   2.1. Where is the work intended to be published?
   2.2. Please list the paper’s authors in the intended authorship order
   2.3. Stage of publication – Not yet submitted / Submitted / Undergoing revision from peer reviewers’ comments / In press

3. For multi-authored work, give full details of your role in the research included in the paper and in the preparation of the paper. (Attach a further sheet if necessary)
   The candidate was involved in study design, marker choice and was responsible for complete characterization of four of the nine housekeeping genes under evaluation. The candidate also participated in data interpretation and assisted in manuscript drafting.

NAME IN FULL (Block Capitals) LOUISA ALEXANDRA MESSENGER
STUDENT ID NO: 223021
CANDIDATE’S SIGNATURE Date: 10/2/15

SUPERVISOR/SENIOR AUTHOR’S SIGNATURE (3 above)
Improving health worldwide www.lshtm.ac.uk
Multilocus Sequence Typing (MLST) for Lineage Assignment and High Resolution Diversity Studies in *Trypanosoma cruzi*

Matthew Yeo1*, Isabel L. Mauricio2, Louisa A. Messenger1, Michael D. Lewis1, Martin S. Llewellyn1, Nidia Acosta1, Tapan Bhattacharya1, Patricio Diosque6, Hernan J. Carrasco5, Michael A. Miles1

1 London School of Hygiene and Tropical Medicine, London, United Kingdom, 2 Instituto de Higiene e Medicina Tropical, Lisboa, Portugal, 3 Departamento de Medicina Tropical, Instituto de Investigaciones en Ciencias de la Salud, Universidad Nacional de Asuncion, Asuncion, Paraguay, 4 Unidad de Epidemiología Molecular (UEM), Instituto de Patología Experimental, Universidad Nacional de Salta, Salta, Argentina, 5 Instituto de Medicina Tropical, Universidad Central de Venezuela, Los Chaguaramos, Caracas, Venezuela

**Abstract**

**Background:** Multilocus sequence typing (MLST) is a powerful and highly discriminatory method for analysing pathogen population structure and epidemiology. *Trypanosoma cruzi*, the protozoan agent of American trypanosomiasis (Chagas disease), has remarkable genetic and ecological diversity. A standardised MLST protocol that is suitable for assignment of *T. cruzi* isolates to genetic lineage and for higher resolution diversity studies has not been developed.

**Methodology/Principal Findings:** We have sequenced and diplotyped nine single copy housekeeping genes and assessed their value as part of a systematic MLST scheme for *T. cruzi*. A minimum panel of four MLST targets (Met-III, RB19, TcGPXII, and DHFR-TS) was shown to provide unambiguous assignment of isolates to the six known *T. cruzi* lineages (Discrete Typing Units, DTUs TcI–TcVI). In addition, we recommend six MLST targets (Met-II, Met-III, RB19, TcMPX, DHFR-TS, and TR) for more in depth diversity studies on the basis that diploid sequence typing (DST) with this expanded panel distinguished 38 out of 39 reference isolates. Phylogenetic analysis implies a subdivision between North and South American TcIV isolates. Single Nucleotide Polymorphism (SNP) data revealed high levels of heterozygosity among DTUs TcI, TcII, TcIII and TcIV, and, for three targets, putative corresponding homozygous and heterozygous loci within DTUs TcI and TcII. Furthermore, individual gene trees gave incongruent topologies at inter- and intra-DTU levels, inconsistent with a model of strict clonality.

**Conclusions/Significance:** We demonstrate the value of systematic MLST diplotyping for describing inter-DTU relationships and for higher resolution diversity studies of *T. cruzi*, including presence of recombination events. The high levels of heterozygosity will facilitate future population genetics analysis based on MLST haplotypes.


**Editor:** Adalgisa Caccone, Yale University, United States of America

**Received** August 17, 2010; **Accepted** May 23, 2011; **Published** June 21, 2011

**Copyright:** © 2011 Yeo et al. This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

**Funding:** This research is funded by the Wellcome Trust (http://www.wellcome.ac.uk/). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

**Competing Interests:** The authors have declared that no competing interests exist.

* E-mail: Matthew.yeo@lshtm.ac.uk

**Introduction**

*Trypanosoma cruzi* is the causative agent of Chagas disease and the most important parasitic infection in Latin America. Approximately 8 million people are thought to be infected [1]. Chagas disease is a zoonosis: *T. cruzi* infects many mammal species [2] and is transmitted to humans primarily by the infected faeces of haematophagous triatomine bugs coming into contact with mucosal membranes or broken skin. Transmission may also be by blood transfusion, congenitally or, rarely, by ingestion of food contaminated by infected triatomine faeces [3].

*T. cruzi* is monophyletic but genetically heterogeneous with at least six phylogenetic lineages (discrete typing units, DTUs) previously designated TcI, and TcIIc–e. A recent review meeting on *T. cruzi* intraspecific nomenclature reached an international consensus that these six DTUs should be renamed (former nomenclature in parenthesis): TcI (TcI), TcII (TcIIb), TcIII (TcIIc), TcIV (TcIIa), TcV (TcIIId) and TcVI (TcIIe), to remove the presumptive sublineage designations within TcII [4]. Here we will adopt this new consensus nomenclature.

TcI and TcII are the most genetically distant groups. The evolutionary ancestry of TcIII and TcIV is presently a debated issue. Based on sequencing of individual nuclear genes Westeriiberger et al. [5] suggested an ancient hybridisation event occurred between TcI and TcII followed by a long period of clonal propagation leading to the extant DTUs TcIII and TcIV. Alternatively, de Freitas et al. [6] suggested that TcIII and TcIV have a separate evolutionary ancestry with mitochondrial sequences that are similar to each other but distinct from both TcI and TcII. Less controversially it is clear, using an array of molecular markers [6,7,8] that TcV and TcVI are hybrid lineages sharing haplotypes from both TcII and TcIII, with both DTUs retaining...
The single-celled parasite Trypanosoma cruzi occurs in animals and insect vectors in the Americas. When transmitted to humans it causes a major public health problem, Chagas disease (American trypanosomiasis). T. cruzi is genetically diverse and currently split into six groups, known as Tcl to TcVI. Multilocus sequence typing (MLST) is a method used for studying the population structure and diversity of pathogens. MLST involves sequencing the DNA of several different genes and comparing the sequences between isolates. MLST has not yet been developed and systematically applied to T. cruzi. He, we sequence nine T. cruzi genes, selecting a panel of four for lineage assignment and six for higher resolution studies of genetic diversity. Our results showed that one of the T. cruzi genetic groups is further subdivided into North and South American subpopulations. Furthermore, comparative analyses of the gene sequences gave new evidence of genetic exchange in T. cruzi. Application of MLST for assigning field isolates of T. cruzi to genetic groups and for detailed investigation of diversity provides a valuable approach to understanding the taxonomy, population structure, genetics, ecology and epidemiology of this important human pathogen.

Multilocus sequencing (MLST) has been applied to T. cruzi. Previous methods for characterisation have included multilocus enzyme electrophoresis [11,12,13,14], PCR amplification of single gene loci, [15] and various PCR based assays [16,17]. Lewis et al. [18] recommended the use of a triple assay comprising SSU rRNA, HSP60 and GPI markers, allowing reliable and low cost typing to DTU level. However, the method is potentially highly discriminatory, when sequence polymorphisms within each housekeeping locus are categorised as distinct alleles. Sequence data for all loci are usually concatenated to produce a unique allelic profile (sequence type, ST). A major advantage of MLST analysis is that sound sequence data are unambiguous and suitable for population structure and epidemiological studies. Results are particularly relevant if easily accessible international databases are constructed such as MLST.net [25] which contains MLST typing schemes and data for a growing number of pathogens.

MLST was originally designed for haploid species, T. cruzi is diploid organism and as such heterozygosity renders MLST analysis more complicated. Heterozygosity from electropherograms can be inferred by a double peak (with two bases) at the same variable bi-allelic site [23]. One consequence of multiple bi-allelic sites is that of ambiguous allelic phase within loci and also ambiguous combinations of alleles across separate loci. However, it is possible for diploid sequence data, (without phase resolution) to be modified, concatenated across multiple loci [26] and applied in distance based phylogenetic methods for lineage assignment (see Materials and Methods). From these data one can also infer evolutionary relatedness, and detect gene mosaics, within or between homoygous gene loci [27]. A MLST approach specific to T. cruzi was first utilised by Machado and Ayala using two nuclear loci [8] in their study of the genetic recombination in natural T. cruzi populations and this is now being expanded by others to additional targets [28]. However the use of multilocus target genes is not recommended due to the possibility of non-identical paralogous copies confounding phylogenetic signals. The repitive nature of the T. cruzi genome has previously hampered the search for suitable single copy targets; at least 50% of the T. cruzi genome consists of tandemly repeated genes [29]. However, the recent publication of a draft chromosomal level assembly by Weatherly et al. [30] has enabled both the determination of copy number and chromosomal distribution of markers to be established with reasonable confidence. A standardised panel of suitable MLST gene loci has yet to be developed for T. cruzi. Here we address this omission by sequencing and assessing 9 nuclear targets and evaluating them with a cohort of reference strains representing the known lineages. We demonstrate the potential of this formalised MLST for describing lineage assignment, describing inter-DTU relationships and for high resolution population genetic analysis of T. cruzi.

Materials and Methods

Isolates
A panel of 39 individual isolates (Table 1) was assembled, consisting of cloned reference strains encompassing all of the known DTUs (Table 1) spanning wide geographical and varied ecological origins. Isolates were previously characterised to DTU level by amplified fragment length polymorphisms (RFLP) in the D7 divergent domain of the 24S rRNA, and restriction fragment length polymorphism (RFLP) in the heat shock protein 60 (HSP60) and GPI genes [18]. Parasites were cultivated in supplemented RPMI liquid medium at 28°C, as described previously [31]. Genomic DNA was prepared from logarithmic phase cultures using DNeasy kits (Qiagen, UK).

Choice of loci
Initially, 11 genes were investigated. This number was subsequently reduced to 9 single copy targets after screening for reliability of PCR amplification. The excluded genes were trypanothione-dependent glyoxalase I (gene ID, Tc00.104705351 0659.240) and cyclophilin (gene ID, Tc00.1047053510947.50). Targets were verified by PCR amplification and sequences submitted for BLAST (blast) analyses hosted at NCBI. Copy number of targets was verified by submission of gene IDs to TriTrypDB 2.2 (http://TriTrypDB.org). The 9 single copy gene fragments amplified for MLST analysis were ascorbate-dependent glutathione reductase-thymidylate synthase (DHFR-TS), dihydrofolate reductase-thymidylate synthase (DHFR-TS), glutathione-dependent peroxidase II (GpxII), mitochondrial peroxidase (TcGpx), trypanothione reductase (TrR), RNA-binding protein-19 (RB19), metacyclin-II (Met-II), metacyclin-III (Met-III) and L17I. The ratio of non-synonymous to synonymous amino acid changes (dN/dS) was calculated according to the Nei-Gojobori method [32] using SNAP software available at http://www.hiv.lanl.gov, [33] to infer relative selection pressures. Genes possessing a dN/dS ratio <1 meet the criteria for stabilising selection for the conservation of
metabolic function. Taylor & Fisher [34] recommended the incorporation of some loci with a dN/dS ratio of >1 in order to obtain sufficient sequence diversity.

Table 1. Cohort of reference clonal isolates representing the six known T. cruzi lineages (DTUs).

<table>
<thead>
<tr>
<th>Strain</th>
<th>DTU</th>
<th>Origin</th>
<th>Host</th>
</tr>
</thead>
<tbody>
<tr>
<td>C8 c11</td>
<td>TcI</td>
<td>La Paz, Bolivia</td>
<td>Triatoma infestans</td>
</tr>
<tr>
<td>X10/1</td>
<td>TcI</td>
<td>Belem, Brazil</td>
<td>Homo sapiens</td>
</tr>
<tr>
<td>JR c4</td>
<td>TcI</td>
<td>Anzoategui, Venezuela</td>
<td>Homo sapiens</td>
</tr>
<tr>
<td>PI (CJ007)</td>
<td>TcI</td>
<td>Carajas, Brazil</td>
<td>Didelphis marsupialis</td>
</tr>
<tr>
<td>PII (CJ005)</td>
<td>TcI</td>
<td>Carajas, Brazil</td>
<td>Unidentified triatomine</td>
</tr>
<tr>
<td>B187 c10</td>
<td>TcI</td>
<td>Pará State, Brazil</td>
<td>Didelphis marsupialis</td>
</tr>
<tr>
<td>SAXP18 c1</td>
<td>TcI</td>
<td>Majes, Peru</td>
<td>Didelphis marsupialis</td>
</tr>
<tr>
<td>92101601P c1</td>
<td>TcI</td>
<td>Georgia, U.S.A.</td>
<td>Didelphis marsupialis</td>
</tr>
<tr>
<td>PII (CJ007)</td>
<td>TcI</td>
<td>Carajas, Brazil</td>
<td>Unidentified triatomine</td>
</tr>
<tr>
<td>92–80 cl2</td>
<td>TcI</td>
<td>Santacruz, Bolivia</td>
<td>Didelphis marsupialis</td>
</tr>
<tr>
<td>PARA4 c3</td>
<td>TcIV</td>
<td>Paraguay</td>
<td>Didelphis marsupialis</td>
</tr>
<tr>
<td>PARA6 c4</td>
<td>TcIV</td>
<td>Paraguay</td>
<td>Didelphis marsupialis</td>
</tr>
<tr>
<td>Sc43 c1</td>
<td>TcIV</td>
<td>Santa Cruz, Bolivia</td>
<td>Didelphis marsupialis</td>
</tr>
<tr>
<td>Chaco2 c3</td>
<td>TcIV</td>
<td>Chaco, Paraguay</td>
<td>Didelphis marsupialis</td>
</tr>
<tr>
<td>VFR1 c1</td>
<td>TcIV</td>
<td>Francia, Chile</td>
<td>Didelphis marsupialis</td>
</tr>
</tbody>
</table>

PCR amplification

Primers and annealing temperatures for PCR amplification are given in Table 2. For DHFR-TS and TR, cyclic amplifications were performed with an initial denaturation step for three minutes at 94°C, followed by 30 amplification cycles (94°C for 30 seconds, 58°C for 1 minute, 72°C for 2 minutes). Annealing temperatures were 55°C for TCAPX, 50°C for TCMPS and 62°C for TRGPXII. Reaction conditions for Met-II, and RB19 were as follows: 3 minutes at 94°C, followed by 30 amplification cycles (94°C for 30 seconds, 53°C for 30 seconds, 72°C for 45 seconds; annealing temperatures were 51°C for Met-III and 36°C for LTI). All reactions had a final ten minute elongation step at 72°C. Each 20 μl total reaction volume contained: 125 ng genomic DNA, 1 μl of each primer (20 μmol/μl), 2 μl dNTPs (2 mM), 0.8 μl (50 mM) MgCl₂, and 5U Taq (BIO-21086, Bioline, UK).

PCR products were purified with QIAQuick PCR purification Kits (Qiagen) or SureClean (Bioline, UK). Bidirectional sequencing was performed with Big Dye Terminator Cycle Sequencing V3.1 (Applied Biosystems) in ABI PRISM 377 DNA Sequencers (Applied Biosystems) according to the manufacturer’s protocol. Gene fragments were sequenced in both directions (5’ and 3’) with the PCR primers described in Table 2. Additionally, in the case of LTI1, Met-II, Met-III, DHFR-TS and TR internal primers were used to obtain full sequence coverage. Sequence data were assembled manually in BioEdit v7.0.9.0 sequence alignment editor software (Ibis Biosciences, USA) and ambiguous peripheral regions of aligned sequences discarded to produce unambiguous partial gene sequences for each isolate. Chromatograms were examined visually in both directions and in most instances the results easily interpreted as heterozygous when two peaks in a chromatogram overlap. Re-sequencing was undertaken if results were ambiguous.

Strain differentiation by diploid sequence typing

The number of sequence types (STs) for each gene fragment was identified from SNP data across the panel of isolates and the discriminatory power (DP, genotypes recovered per unique isolate tested) for each gene determined. STs were conjoined across gene fragments in order to identify a diploid sequence type (DST) and to assess overall discriminatory power (DP). DSTs were applied to eBURST software to infer evolutionary relationships and founders. The eBURST algorithm (http://euburst.mlst.net) identifies related sequences and predicts a founding genotype (based on the most overrepresented genotype) with variants identified depending on the number of different loci.

Phylogenetic analysis was performed on modified sequences using the methodology described by Tavanti et al. [26] which renders diplotypic data suitable for use in MEGA analysis [35] to produce distance based dendrograms. Briefly, variable loci between isolates can be considered either homozygous or heterozygous, assuming that T. cruzi is minimally diploid. For example, a homozygous variable locus scored as A (adenine) was modified by duplication to AA, and a heterozygous locus, for example Y (C or T, in accordance with IUPAC nomenclature), scored as CT, effectively creating a difference matrix across the panel. Phylogenetically the data were examined in two different ways. Firstly sequences were examined at the level of individual genes by generating neighbor-joining trees (MEGA v4.0.2) in order to assess the “usefulness” of gene fragments in the context of lineage assignment, intralineage resolution and associated bootstrap values. Secondly, sequence data were concatenated across multiple gene fragments to produce MLST neighbor-joining trees.
Table 2. Details of gene targets.

<table>
<thead>
<tr>
<th>Gene</th>
<th>Gene ID</th>
<th>Chromosome Number</th>
<th>Primer Sequence (5' → 3')</th>
<th>Annealing Temp. (C)</th>
<th>Amplicon Size (bp)</th>
<th>Sequence Start 5'</th>
<th>Sequence Start 3'</th>
<th>Fragment Length (bp)</th>
</tr>
</thead>
<tbody>
<tr>
<td>LYT1</td>
<td>Tc00.1047053508045.40</td>
<td>22</td>
<td>CAACCTGCTCCCTTTGGCTCTG (20)</td>
<td>56</td>
<td>804</td>
<td>TTTTGAC</td>
<td>GTTGTTGA</td>
<td>691</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>GGTGGTGTTGCGGCTTGTG (20)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>AGAAGGCTGTTCGGTGAAG (20)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Met II</td>
<td>Tc00.1047053510889.28</td>
<td>6</td>
<td>TCATCTGCACCGAGTGCTGTC (20)</td>
<td>53</td>
<td>702</td>
<td>CATTTCGA</td>
<td>TTTTGCCA</td>
<td>390</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>CTCATAGCGGTGACGAACA (20)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>TGTCACACATTGCTGTGTT (20)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Met III</td>
<td>Tc00.1047053510943.44</td>
<td>36</td>
<td>GGTGCTCCAAGACATTAAGA (20)</td>
<td>51</td>
<td>824</td>
<td>TTCAATCCG</td>
<td>TTTCTTTG</td>
<td>619</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>CCCCCCTCCTTGGCAAATAT (20)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>CACATGAGTCTGTGTTGCT (20)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>RbI19</td>
<td>Tc00.1047053507505.60</td>
<td>29</td>
<td>GCCTACACCCAGAGGATGCAA (20)</td>
<td>53</td>
<td>408</td>
<td>GTGTGCAG</td>
<td>CCCAGACT</td>
<td>350</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>TTCTCCATTCCCCAGACTCTG (20)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>TcAPX</td>
<td>Tc00.1047053506193.60</td>
<td>36</td>
<td>GGGAAGCCTTACCTTGGCTTATTTTGC (30)</td>
<td>55</td>
<td>1500</td>
<td>GCGAGTCCG</td>
<td>GGGCCCGCG</td>
<td>799</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>GGGGATCCATGGCTTTTGGTTTGTG (30)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>GGTGCA (30)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>TcMPX</td>
<td>Tc00.1047053509499.14</td>
<td>22</td>
<td>ATGTTCGGCTGATGGCC (18)</td>
<td>50</td>
<td>678</td>
<td>TACATGGA</td>
<td>GCACGCCT</td>
<td>505</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>TCCTTCTTCTCAATATCTTC (21)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>DHFR-TS</td>
<td>Tc00.1047053505935.90</td>
<td>27</td>
<td>CGGCCTTAAGATCCGATGCCG (22)</td>
<td>58</td>
<td>1475</td>
<td>GGGGGAAGA</td>
<td>CACGCTCT</td>
<td>715</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>CGATAGTCAATGACCTCCATGTC (24)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>AACCCTCTCGGCTCAGTG (19)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>GCCACCGTCTTG (21)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>TcGPXII</td>
<td>Tc00.1047053511019.99</td>
<td>35</td>
<td>GGCGAAGCGCAAGCGCTGCGC (20)</td>
<td>62</td>
<td>487</td>
<td>TTGTGGGC</td>
<td>ATCCGGCC</td>
<td>437</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>TCATCCCCGCTGGCCTGCC (21)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7R</td>
<td>Tc00.1047053503555.30</td>
<td>37</td>
<td>ACTGAGGTGCTTGAGGAAAGG (21)</td>
<td>58</td>
<td>1290</td>
<td>TGCTCAATG</td>
<td>TACGAAAG</td>
<td>602</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>GGTGTGCACAACCTTGGTG (22)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>CGAATGARGCATTTTACCTG (20)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>TACTCTGCACCTGGACACCCAC (22)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Analysis of recombination at the level of individual genes was applied to isolates with unambiguous phase applied through the software package RDP5 [36] incorporating the following methods: RDP [37], Bootscanning [38], GENECONV [39], Maximum Chi Square method [40,41], the Chimaera method [40], the Sister Scanning Method [42], the SSEQ method [43], the Reticulate compatibility matrix method [44] and the TOPAL DSS method [45,46].

Results

Chromosomal localisation of markers

Gene IDs were confirmed by BLAST searches and submitted to the TriTrypDB 1.3 (http://TriTrypDB.org) to identify chromosomal location. Locations for each of the the genes under study are shown in Table 2. Two pairs of genes were physically linked: genes LYT1 (367579–369237) and TcMPX (P:120685–121365) on chromosome 22, and MET-III (P:945156–945731) and TcAPX (P:1142201–1143187) on chromosome 36. The remaining genes were on independent chromosomes. The ratios of nonsynonymous to synonymous amino acid substitutions, resulting from sequence polymorphisms, were 1.0 or less for 7 genes, two genes possessed dN/dS ratios of above 1 (Table 3).

Nucleotide polymorphisms and amino acid changes

The 304 nucleotide polymorphisms among the nine sequenced fragments resulted in 121 nonsynonymous changes in amino acids encoded by sequence-variable triplets. Up to three different amino acids were present in different isolates at the same locus across the panel. A deletion of 3 nucleotides (AAA = Lysine at position 179–181) was present in C8, X10/1 and SAXP18 (TcI) within LYT1. Within TcAPX, a single non-synonymous nucleotide polymorphism resulted in a conservative amino acid change between an aspartic acid residue and a glutamic acid residue was evident (position 21). All TcV representatives possessed glutamic acid but aspartic acid residue. A single non-synonymous nucleotide polymorphism resulted in a conservative amino acid change between an aspartic acid residue at the corresponding position.

Discriminatory power of MLST targets by diploid sequence typing

Diploid sequence typing using 9 genes was able to discriminate 38 of 39 isolates (DP 0.97, Table 4). Importantly, using only 6 genes (Met-II, Met-III, RB19, TcMPX, DHFR-TS and TR) the same number of DSTs could be identified. Only two isolates, Sc43 and Vinch101, shared a DST (29). Both are positioned within the DTU TcV which is known to possess a particularly homogenous population structure by microsatellite analysis [10]. All other isolates, notably even those those within the relatively homogenous clade TcVI, produced a unique isolate specific DST. Discriminatory power was determined for each of the 9 gene fragments (Table 4), in decreasing power, as follows: Met-II, Met-III and RB19 (0.59), LYT1 (0.56), TR (0.54), DHFR-TS and TcGPXII (0.49), TcAPX (0.41), and TcAPX (0.38). No single gene was able to distinguish all 39 reference strains. Met-II, Met-III and RB19, taken individually, were the most discriminatory genes, all three identifying 24 separate genotypes from the panel of 39. TcAPX was the least resolutive marker distinguishing 15 of 39 isolates (DP 0.38). Table 4 shows in detail the number of STs (sequence types) that each individual gene fragment resolved, and also the derived DSTs obtained from the concatenation of all 9 STs for each isolate. Table 4 also indicates the reduced panel of 6 genes required to obtain the same 38 DSTs. eBURST analysis of the genotypes and DSTs for 39 T. cruzi isolates from the panel revealed one cluster of 7 DSTs within TcV as the only related set to emerge from this analysis. Vinch101 (Lanari, Chile) and Sc43 (Santa Cruz, Bolivia), both DST 29 (Table 4) are the predicted genetic founders of isolates of all other TcV isolates. The remaining isolates across other DTUs appeared as unrelated singletons (isolates that do not belong to any cluster). Overall the results reveal that diploid sequence typing using just 6 genes is highly discriminatory.

Intra DTU diversity

Table 5 describes the levels of diversity seen in each gene fragment represented as the number of variable sites per DTU (VS), the number of genotypes differentiated per DTU (GT) and the discriminatory power for each gene fragment for each DTU (P:1142201–1143187) on chromosome 36. The remaining genes were on independent chromosomes. The ratios of nonsynonymous to synonymous amino acid substitutions, resulting from sequence polymorphisms, were 1.0 or less for 7 genes, two genes possessed dN/dS ratios of above 1 (Table 3).

Table 3. Properties of nine T. cruzi MLST targets.

<table>
<thead>
<tr>
<th>Gene Fragment</th>
<th>No. Of Polymorphic Sites</th>
<th>No. Of Genotypes</th>
<th>No. Of Genotypes/Polymorphism (Typing Efficiency)</th>
<th>Ratio Of Nonsynonymous To Synonymous Changes</th>
</tr>
</thead>
<tbody>
<tr>
<td>LYT1</td>
<td>47</td>
<td>22</td>
<td>0.47</td>
<td>0.833</td>
</tr>
<tr>
<td>Met-II</td>
<td>51</td>
<td>24</td>
<td>0.47</td>
<td>0.880</td>
</tr>
<tr>
<td>Met-III</td>
<td>50</td>
<td>24</td>
<td>0.48</td>
<td>0.440</td>
</tr>
<tr>
<td>RB19</td>
<td>27</td>
<td>24</td>
<td>0.89</td>
<td>0.129</td>
</tr>
<tr>
<td>TcAPX</td>
<td>27</td>
<td>14</td>
<td>0.52</td>
<td>2.04</td>
</tr>
<tr>
<td>TcMPX</td>
<td>15</td>
<td>16</td>
<td>1.06</td>
<td>0.061</td>
</tr>
<tr>
<td>DHFR-TS</td>
<td>32</td>
<td>19</td>
<td>0.59</td>
<td>0.088</td>
</tr>
<tr>
<td>TcGPXII</td>
<td>27</td>
<td>18</td>
<td>0.66</td>
<td>0.502</td>
</tr>
<tr>
<td>TR</td>
<td>28</td>
<td>21</td>
<td>0.75</td>
<td>1.964</td>
</tr>
</tbody>
</table>

doi:10.1371/journal.pntd.0001049.t003
structures with low intralineage diversity. Nevertheless, TcAPX resolved three of 5 isolates within TcV (DP 0.43) and TcMPX 5 of 6 (DP 0.83) isolates within DTU VI.

Phylogenies and incongruence

In addition to the derivation of DSTs, neighbor-joining trees for individual and concatenated gene fragments were constructed, representatives of which are shown in Figures 1 and 2. Concatenation using all of the original 9 gene fragments generated phylogenies with the expected DTU assignments for all 39 isolates with much higher bootstrap values than for individual genes (supporting information, Figure S1); there were no unexpected outliers. However, bootstrap support distinguishing TcV and VI was low (21%), but to distinguish between these DTUs concatenation of just two genes (DHFR-TS and LYT1 sequences) generated robust lineage assignment, and also two separate clusters within

### Table 4. Sequence types (STs) and diploid sequence types (DSTs) for nine gene fragments.

<table>
<thead>
<tr>
<th>Strain</th>
<th>DTU</th>
<th>DST</th>
<th>LYT1 (ST)</th>
<th>Met-II*</th>
<th>Met-III*</th>
<th>RB19*</th>
<th>TcAPX</th>
<th>TcMPX*</th>
<th>DHFR-TS*</th>
<th>TcGPXII</th>
<th>TR*</th>
</tr>
</thead>
<tbody>
<tr>
<td>C8 cl1</td>
<td>Tcl</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>X10/1</td>
<td>Tcl</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>JR cl4</td>
<td>Tcl</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>2</td>
<td>3</td>
<td>3</td>
<td>1</td>
<td>1</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>PI (CJO07)</td>
<td>Tcl</td>
<td>4</td>
<td>4</td>
<td>1</td>
<td>1</td>
<td>4</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>PI (CJO05)</td>
<td>Tcl</td>
<td>5</td>
<td>5</td>
<td>4</td>
<td>3</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>3</td>
</tr>
<tr>
<td>B187 cl10</td>
<td>Tcl</td>
<td>6</td>
<td>6</td>
<td>4</td>
<td>1</td>
<td>5</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>SAXP18 cl1</td>
<td>Tcl</td>
<td>7</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>6</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>5</td>
</tr>
<tr>
<td>92101601P cl1</td>
<td>Tcl</td>
<td>8</td>
<td>7</td>
<td>1</td>
<td>4</td>
<td>7</td>
<td>2</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>6</td>
</tr>
<tr>
<td>Eim cl3</td>
<td>Tcl</td>
<td>9</td>
<td>8</td>
<td>5</td>
<td>5</td>
<td>8</td>
<td>4</td>
<td>2</td>
<td>3</td>
<td>5</td>
<td>7</td>
</tr>
<tr>
<td>Pot7a cl1</td>
<td>Tcl</td>
<td>10</td>
<td>8</td>
<td>5</td>
<td>6</td>
<td>8</td>
<td>4</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>7</td>
</tr>
<tr>
<td>Pot7b cl2</td>
<td>Tcl</td>
<td>11</td>
<td>8</td>
<td>5</td>
<td>5</td>
<td>8</td>
<td>4</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>8</td>
</tr>
<tr>
<td>Tu18 cl2</td>
<td>Tcl</td>
<td>12</td>
<td>9</td>
<td>6</td>
<td>7</td>
<td>9</td>
<td>5</td>
<td>2</td>
<td>5</td>
<td>5</td>
<td>8</td>
</tr>
<tr>
<td>Chaco23 col4</td>
<td>Tcl</td>
<td>13</td>
<td>10</td>
<td>7</td>
<td>8</td>
<td>9</td>
<td>4</td>
<td>2</td>
<td>6</td>
<td>6</td>
<td>8</td>
</tr>
<tr>
<td>M5631 cl5</td>
<td>Tcl</td>
<td>14</td>
<td>11</td>
<td>8</td>
<td>9</td>
<td>10</td>
<td>6</td>
<td>4</td>
<td>7</td>
<td>7</td>
<td>9</td>
</tr>
<tr>
<td>M6421 cl6</td>
<td>Tcl</td>
<td>15</td>
<td>11</td>
<td>9</td>
<td>10</td>
<td>11</td>
<td>7</td>
<td>4</td>
<td>7</td>
<td>8</td>
<td>10</td>
</tr>
<tr>
<td>ARMA18 cl3</td>
<td>Tcl</td>
<td>16</td>
<td>12</td>
<td>10</td>
<td>10</td>
<td>12</td>
<td>6</td>
<td>5</td>
<td>7</td>
<td>8</td>
<td>11</td>
</tr>
<tr>
<td>ARMA13 cl1</td>
<td>Tcl</td>
<td>17</td>
<td>12</td>
<td>9</td>
<td>11</td>
<td>13</td>
<td>6</td>
<td>5</td>
<td>7</td>
<td>9</td>
<td>11</td>
</tr>
<tr>
<td>JA2 cl2</td>
<td>Tcl</td>
<td>18</td>
<td>13</td>
<td>11</td>
<td>12</td>
<td>14</td>
<td>6</td>
<td>4</td>
<td>8</td>
<td>10</td>
<td>10</td>
</tr>
<tr>
<td>CM25 cl2</td>
<td>Tcl</td>
<td>19</td>
<td>14</td>
<td>12</td>
<td>13</td>
<td>15</td>
<td>7</td>
<td>4</td>
<td>7</td>
<td>11</td>
<td>12</td>
</tr>
<tr>
<td>85/847 cl2</td>
<td>Tcl</td>
<td>20</td>
<td>15</td>
<td>11</td>
<td>14</td>
<td>16</td>
<td>6</td>
<td>4</td>
<td>9</td>
<td>12</td>
<td>13</td>
</tr>
<tr>
<td>SABP19 cl5</td>
<td>Tcl</td>
<td>21</td>
<td>16</td>
<td>13</td>
<td>15</td>
<td>17</td>
<td>6</td>
<td>6</td>
<td>7</td>
<td>7</td>
<td>11</td>
</tr>
<tr>
<td>StCl10 cl1</td>
<td>Tcl</td>
<td>22</td>
<td>17</td>
<td>14</td>
<td>16</td>
<td>18</td>
<td>8</td>
<td>7</td>
<td>10</td>
<td>13</td>
<td>14</td>
</tr>
<tr>
<td>92122102R</td>
<td>Tcl</td>
<td>23</td>
<td>17</td>
<td>15</td>
<td>17</td>
<td>18</td>
<td>8</td>
<td>8</td>
<td>11</td>
<td>14</td>
<td>15</td>
</tr>
<tr>
<td>1R26</td>
<td>Tcl</td>
<td>24</td>
<td>18</td>
<td>16</td>
<td>18</td>
<td>19</td>
<td>9</td>
<td>9</td>
<td>12</td>
<td>15</td>
<td>16</td>
</tr>
<tr>
<td>Can III cl1</td>
<td>Tcl</td>
<td>25</td>
<td>19</td>
<td>17</td>
<td>19</td>
<td>20</td>
<td>10</td>
<td>10</td>
<td>13</td>
<td>16</td>
<td>17</td>
</tr>
<tr>
<td>Samirn c1</td>
<td>Tcl</td>
<td>26</td>
<td>20</td>
<td>18</td>
<td>20</td>
<td>20</td>
<td>11</td>
<td>11</td>
<td>14</td>
<td>17</td>
<td>18</td>
</tr>
<tr>
<td>PARA4 c3</td>
<td>Tcl</td>
<td>27</td>
<td>21</td>
<td>19</td>
<td>21</td>
<td>21</td>
<td>12</td>
<td>11</td>
<td>15</td>
<td>18</td>
<td>19</td>
</tr>
<tr>
<td>PARA6 c4</td>
<td>Tcl</td>
<td>28</td>
<td>21</td>
<td>20</td>
<td>22</td>
<td>21</td>
<td>12</td>
<td>11</td>
<td>15</td>
<td>18</td>
<td>19</td>
</tr>
<tr>
<td>S433 cl1</td>
<td>Tcl</td>
<td>29</td>
<td>21</td>
<td>19</td>
<td>22</td>
<td>21</td>
<td>13</td>
<td>11</td>
<td>16</td>
<td>18</td>
<td>19</td>
</tr>
<tr>
<td>92–80 cl2</td>
<td>Tcl</td>
<td>30</td>
<td>21</td>
<td>19</td>
<td>22</td>
<td>21</td>
<td>13</td>
<td>11</td>
<td>15</td>
<td>18</td>
<td>19</td>
</tr>
<tr>
<td>Chaco2 cl3</td>
<td>Tcl</td>
<td>31</td>
<td>21</td>
<td>19</td>
<td>22</td>
<td>21</td>
<td>13</td>
<td>11</td>
<td>16</td>
<td>18</td>
<td>20</td>
</tr>
<tr>
<td>VINc101 cl1</td>
<td>Tcl</td>
<td>29</td>
<td>21</td>
<td>19</td>
<td>22</td>
<td>21</td>
<td>13</td>
<td>11</td>
<td>16</td>
<td>18</td>
<td>19</td>
</tr>
<tr>
<td>FAH179 cl5</td>
<td>Tcl</td>
<td>32</td>
<td>21</td>
<td>19</td>
<td>22</td>
<td>21</td>
<td>14</td>
<td>12</td>
<td>16</td>
<td>18</td>
<td>19</td>
</tr>
<tr>
<td>CL Brener</td>
<td>Tcl</td>
<td>33</td>
<td>22</td>
<td>21</td>
<td>23</td>
<td>22</td>
<td>13</td>
<td>13</td>
<td>17</td>
<td>18</td>
<td>19</td>
</tr>
<tr>
<td>Tula cl2</td>
<td>Tcl</td>
<td>34</td>
<td>22</td>
<td>21</td>
<td>24</td>
<td>23</td>
<td>13</td>
<td>14</td>
<td>18</td>
<td>18</td>
<td>19</td>
</tr>
<tr>
<td>P251 cl7</td>
<td>Tcl</td>
<td>35</td>
<td>22</td>
<td>21</td>
<td>24</td>
<td>22</td>
<td>13</td>
<td>8</td>
<td>18</td>
<td>18</td>
<td>19</td>
</tr>
<tr>
<td>EPV20–1 cl1</td>
<td>Tcl</td>
<td>36</td>
<td>22</td>
<td>22</td>
<td>23</td>
<td>22</td>
<td>13</td>
<td>15</td>
<td>18</td>
<td>18</td>
<td>19</td>
</tr>
<tr>
<td>LHVA cl4</td>
<td>Tcl</td>
<td>37</td>
<td>22</td>
<td>23</td>
<td>23</td>
<td>22</td>
<td>14</td>
<td>8</td>
<td>18</td>
<td>18</td>
<td>19</td>
</tr>
<tr>
<td>VFRA1 c1</td>
<td>Tcl</td>
<td>38</td>
<td>22</td>
<td>24</td>
<td>23</td>
<td>24</td>
<td>15</td>
<td>16</td>
<td>19</td>
<td>18</td>
<td>21</td>
</tr>
</tbody>
</table>

*Indicate six genes required to identify the same number of DSTs as the full panel of 9 genes.

doi:10.1371/journal.pntd.0001049.t004
Table 5. Intra-lineage diversity.

<table>
<thead>
<tr>
<th>SNP per DTU</th>
<th>LYT1</th>
<th>Met-II</th>
<th>Met-III</th>
<th>RB19</th>
<th>TcAPX</th>
<th>TcMPX</th>
<th>DHFR-TS</th>
<th>TcGPXII</th>
<th>TR</th>
</tr>
</thead>
<tbody>
<tr>
<td>Isolates</td>
<td>VS</td>
<td>GT</td>
<td>DP</td>
<td>VS</td>
<td>GT</td>
<td>DP</td>
<td>VS</td>
<td>GT</td>
<td>DP</td>
</tr>
<tr>
<td>TcI</td>
<td>8</td>
<td>8</td>
<td>7</td>
<td>0.88</td>
<td>10</td>
<td>4</td>
<td>0.50</td>
<td>6</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>TcII</td>
<td>5</td>
<td>6</td>
<td>3</td>
<td>0.60</td>
<td>3</td>
<td>3</td>
<td>0.60</td>
<td>5</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>TcIII</td>
<td>8</td>
<td>7</td>
<td>6</td>
<td>0.75</td>
<td>7</td>
<td>6</td>
<td>0.75</td>
<td>12</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>TcIV</td>
<td>5</td>
<td>13</td>
<td>4</td>
<td>0.80</td>
<td>10</td>
<td>5</td>
<td>1.00</td>
<td>10</td>
<td>5</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>TcV</td>
<td>7</td>
<td>0</td>
<td>1</td>
<td>0.14</td>
<td>18</td>
<td>2</td>
<td>0.29</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>TcVI</td>
<td>6</td>
<td>0</td>
<td>1</td>
<td>0.17</td>
<td>14</td>
<td>4</td>
<td>0.67</td>
<td>20</td>
<td>2</td>
</tr>
</tbody>
</table>

Abbreviations in each of the gene fragment columns are as follows: VS = No variable sites per DTU, GT = Number of genotypes per DTU, DP = Discriminatory power. Numbers in bold indicate highest DP values per DTU. *All SNP variation observed within the DTU originated from a single isolate.

Choice of loci for lineage assignment

The minimum number of genes required for phylogenetic assignment to DTU level was investigated according to the following criteria. Firstly, the minimum number of genes required to produce phylogenetically robust bootstrap support. Secondly, to choose those genes that produced the most consistent unambiguous sequences. Thirdly, to choose genes requiring the minimum number of internal primers. Using these criteria a minimum combination of 4 genes (Met-III, RB19, TcGPXII and DHFR-TS) enabled all reference isolates to be assigned to DTU level. Specifically, concatenated sequences of Met-III, RB19, and TcGPXII assigned isolates to DTUs TcI-TcIV (Figure 2A). Separately, DHFR-TS (the only gene requiring internal primers) distinguished the DTUs TcV and TcVI (Figure 2B, insert). Internal primers for Met-III, RB19 and TcGPXII were not required after initial optimisation (Table 2).

Intralineage phylogenies

Generally, bootstrap support within DTUs using concatenated sequences was lower than between DTUs (Figure 1) which is expected in data originating from gene fragments with non-contiguous topologies at the intralineage level. In most instances intralineage topology of individual gene dendrograms was generally low and/or not contiguous across loci (Figure 1). However, certain observations deserve closer scrutiny. Importantly, two distinct clusters within TcIV separated isolates from North America and South America. This genetic partitioning was detected in 8 out of 9 gene individual targets in the present study, indicating a robust clades subdivision within TcIV. Interestingly, two TcI isolates SAXP18 (Peru, Didelphis marsupialis) and G8 (Bolivia, Triatoma infestans) were isolated from silvatic and domestic sources respectively, and are genetically similar suggesting present or past overlapping transmission cycles. Relatvely high bootstrap support for isolates of the TcII clade did not reveal evidence of geographical clustering and genetically similar isolates originated...
from distant localities. Within TcIII at least two clusters were seen (in concatenated data); silvatic isolates (from *Dasypus novemcinctus*) in the Paraguayan Chaco clustered with a domestic strain (SABP19) isolated from a domestic *T. infestans* in Peru. The 8 TcIII isolates included in our panel showed considerable diversity. However, numbers are insufficient to examine population substructuring.

Intralineage recombination

Analysis for the presence of mosaic alleles, at the level of individual genetic loci was performed using RDP [37], applied to a total of 240 representative sequences of known allelic phase (single SNP heterozygous and homozygous sequence profiles) encompassing DTUs TcI, TcII, TcIII and TcIV across the 9 genetic loci. We found no evidence of allelic mosaics within individual genes. However, diplotypic SNP data revealed for three genes (*LYT1*, *TcGPXII* and *TcMPX*) that putative donor homozygous SNP profiles and the corresponding heterozygous profiles were present as shown in Figure 3. Specifically, within *LYT1* the heterozygous isolate (PII) and putative donors (B187 and PI) were identified within TcI. For *TcGPXII*, 2 heterozygous isolates (ARMA18 and M6421) possess SNP profiles of the putative donor isolates CM25 and 85/847 (TcIII). Lastly, for *TcMPX*, SABP19 was heterozygous at a single locus with potential donors in the same DTU (TcIII). In each instance individual isolates containing heterozygous and donor SNPs were present only in a single gene and were not contiguous across loci.

SNP data and loss of heterozygosity

Our panel of isolates included 13 representatives of DTUs V and VI, which are known to be genetic hybrids of TcII and TcIII [8].
Across all 9 loci the expected allelic heterozygous profile was observed in most individual isolates. However, at four gene loci complete LOH was detected for individual isolates within the hybrid lineages. The observations are as follows: within Met-III, two TcVI isolates (Tula cl2 and P251 cl7) possessed a TcII-like allele but not a TcIII-like allele as did an individual TcV isolate for Met-II (PARA6 cl4). In contrast a single TcVI isolate at the locus RB19 (VFRA cl1) possessed only the TcIII-like allele. At the TcMPX locus, all TcV isolates (with the exception of PAH179) appeared to have lost the TcIII-like allele and 2 representatives of TcVI (P251 and LHVA) lost the TcII-like allele; the remaining 4 TcVI isolates retained heterozygosity. LOH among individual isolates in genetic loci could be clearly visualised when dendrograms were constructed, for example, hybrid isolates PARA6 (TcV) and EPV20-1(TcVI) for Met-II clustered within TcII (Figure 1). There was no continuity of allelic loss across different genes for individual isolates. For example, the two aforementioned isolates exhibiting homozygosity in Met-III were heterozygous in TcAPX, which are linked on chromosome 36. It is therefore evident that heterozygosity has been maintained elsewhere on the same chromosome. Taken together the results indicate a return to homozygosity in 6.24% of TcV and TcIV (combined) isolates when considering the four affected gene fragments.

Discussion

Discriminatory power and diploid sequence typing

Diploid sequence typing of 39 reference isolates indicates that together the nine genes under study were highly discriminatory, 38 of 39 reference isolates generated a unique DST. It was apparent that the number of STs identified for each gene fragment varied considerably (Table 5) despite the fact that all but two of the genes (TcAPX and TR) were under stabilising selection. However, by using just 6 of the original panel of nine genes the same 38 DSTs could be identified. The reduced panel consisted of those genes with the highest overall DPs (Met-II, Met-III and RB19) with additional genes (TcAPX, DHFR-TS and TR) resolving further STs for the genetically homogenous DTUs TcV and TcVI. The reduction in the number of genes would represent a considerable saving in sequencing effort if applied to new samples. Furthermore, the reduced panel consists primarily of genes that are easily amplified and sequenced. The derived STs and DSTs applied to eBURST software, designed specifically for MLST data to infer evolutionary relationships and genetic founders, identified one cluster encompassing all of the TcV reference isolates. The inferred founder of this group (DST29) was found in 2 isolates (Vinch101 and Sc43). Both of these isolates were originally isolated.
from domestic *T. infestans* from Chile and Bolivia respectively. The ability to differentiate isolates within TcV is an important development, as this DTU is found throughout the Southern Cone countries in domestic cycles. Currently other methods to demonstrate diversity within TcV have been lacking. The high number of singletons observed in remaining DTUs was surprising, although this is probably a consequence of reference isolates spanning diverse spatial and ecological origins and diversity found within our panel. It is likely that high intensity localised sampling would reveal more easily genetic relationships between isolates. High numbers of singletons are often observed in other MLST typing schemes (for example *Candida* species) when applied to small datasets. Furthermore distance based dendrograms have been shown to correlate well with clonal clusters generated by eBURST when applied to large numbers [27]. It is also worth noting that high rates of singletons are typical of populations with a high rate of recombination relative to mutation [27]. Although eBURST is an established software for diplotyping and haplotyping pathogens [47,48] until analysis has been applied to a larger number of sympatric *T. cruzi* isolates it is difficult to evaluate the usefulness of the analysis as a tool for predicting founders and clusters in DTUs other than TcV, and other approaches should be considered in parallel.

**Choice of loci for lineage assignment**

Sequence data concatenated across all 9 genes produced robust phylogenetic assignment for all DTUs except for TcV and TcVI, which could be resolved using two concatenated genes (*DHFRTS* and *LYT1* sequences). Importantly, this number could subsequently be reduced to a combination of just 4 genes (*MtiII*, *RB19*, *TcGPXII* and *DHFRTS*) while maintaining high bootstrap support. The first three aforementioned gene sequences were concatenated to differentiate DTUs TcI-TcIV, with *DHFRTS* dendrograms separately able to distinguish DTUs TcV and VI. Notably concatenated data implies a cladistic subdivision within TcIV, separating North American from South American isolates. Bootstrap support is a requirement for any MLST scheme, but of equal importance are reliable PCR amplifications which generate unambiguous sequences in both forward and reverse directions. All four targets meet this criterion. *MtiII*, *RB19* and *TcGPXII* do not require internal primers, although some were used initially in
the optimisation process, and generate small amplicons (824 bp, 408 bp and 487 bp respectively). DHFR-TS alone requires the use of internal sequencing primers. LYT1 was considered to be an alternative candidate, also able to distinguish TcV and TcVI, but it was technically more difficult to obtain sequences in the forward direction despite multiple attempts at optimisation.

**Intralineage recombination**

SNP data revealed that for three genes (LYT1, TcGPXIII and TcAPX) putative donor homozygous SNP profiles and the corresponding heterozygous profiles, were present within TcI (a single heterozygous isolate) and TcIII (1 heterozygous isolate) one tentative explanation of the presence of heterozygous SNPs is that of a recombination signature within these DTUs. In the present paper examination 240 sequences (of known allelic phase) using an array of recombination detection algorithms did not detect evidence of allelic mosaics. This result is not unexpected as diployping is not as sensitive as haplotyping for detecting mosaic’s or investigating the sexual reproduction in evolutionary history [50]. Future recombination analysis will include more refined phase resolution of sequences heterozygous at two or more loci, typically not required for diployping MLST typing schemes. However, haplotypes can be derived from current targets reconstructed by the program PHASE [51], or more recently fastPHASE [http://depts.washington.edu/uwc4c/express-licenses/assets/fastphase/], by cloning, or by allelic specific PCR [52]. In a sexually reproducing population the frequencies of genotypes (defined by haplotypes) should be in Hardy-Weinberg equilibrium. New software to analyse haplotypic data include those incorporating Bayesian methodologies; Structure [53] and Beast [54] reconstruct phylogenies with epidemiologically and evolutionary informative results.

**Loss of heterozygosity**

LOH observed in 4 gene fragments affecting the hybrid lineages TcV and TcIV has potentially significant consequences for MLST and lineage assignment. For example, TcMPX hybrid isolates (TcV) would be incorrectly assigned if judgment was based on a single locus due to LOH (Figure 1). There was no continuity of allelic loss across different genes for individual isolates, linked single locus due to LOH (Figure 1). There was no continuity of allelic loss across different genes for individual isolates, linked single locus due to LOH (Figure 1). There was no continuity of allelic loss across different genes for individual isolates, linked single locus due to LOH (Figure 1). Consequently this LOH event. Discriminating between recombination, gene conversion, duplication or mitotic recombination requires analysis at the haplotypic level and should be possible by MLST, particularly in light of the considerable heterozygosity found in a T. cruzi (in particular DTUs TcI, TcIII, and TcIV). Use of a larger number of sympatric isolates and gene targets evenly distributed along those chromosomes where LOH has been previously identified (chromosomes 6, 36, 29, 22) would be a suitable approach, and one that has been exploited in C. albicans [55]. Such apparent random loss is consistent with gene conversion. However, within TcAPX all TcIV isolates (with the exception of PAH179) possessed only a TcII like allele suggesting clonal expansion subsequent to this LOH event. The authors thank Christian Barnabé and Michael Tibayrenc, who kindly provided T. cruzi strains. Also thanks to Urvish Mistry for his constructive comments.

**Supporting information**

**Figure S1** Concatenation and lineage assignment. Unrooted neighbor-joining diploypite tree showing p-distance for 9 concatenated gene fragments (A). Concatenation of LYT1 and DHFR-TS discriminate between DTUs V and VI (inset). Concatenated diploypite tree using a reduced panel of 4 gene fragments (B). Concatenation of Met-III, RB19, and TcGPXIII assign isolates to DTUs TcI-TcIV. DHFR-TS differentiates DTUs V and VI (B inset). Found at: doi:10.1371/journal.pntd.0001049.s001 (0.41 MB DOC)

**Acknowledgments**

The authors thank Christian Barnabé and Michael Tibayrenc, who kindly provided T. cruzi strains. Also thanks to Urvish Mistry for his constructive comments.

**Author Contributions**

Conceived and designed the experiments: MY MAM. Performed the experiments: MY LAM. Analyzed the data: MY LAM. Contributed reagents/materials/analysis tools: MY LAM MDM MSL NA TB PD HJC. Wrote the paper: MY MAM. Input into design of experiments and technical advice: ILM.

**References**


3.3.2 Formalization of a standardized MLST scheme for *T. cruzi*

To formalize a standardized MLST scheme for *T. cruzi*, a subset of the most discriminatory genetic markers from two published schemes (Yeo *et al.*, 2011 and Lauthier *et al.*, 2012) were evaluated across a modified panel of 25 *T. cruzi* reference strains, shared by both research groups. A total of thirteen nuclear housekeeping genes were assessed, including three fragments previously examined by Yeo *et al.* (*Met-II, RB19* and *TcMPX*), to determine the optimum combination of loci for: (i) robust bootstrap-supported DTU-level assignment; (ii) DTU monophyly; and (iii) detection of intra-DTU genetic diversity.

This study is reported in full below by Diosque *et al.*, 2014.

In summary:

- The optimum combination of MLST genes consisted of seven loci (*GPI, HMCOAR, LAP1, RB19, RHO1, SODB* and *TcMPX*), which discriminated between all reference strains and separated all DTUs as monophyletic clades (DP=1.0).
- A reduced panel of four MLST targets (*GPI, HMCOAR, RHO1* and *TcMPX*) can be used to assign the majority of isolates to DTU-level (19/25 DSTs; DP=0.76).
- PCR reproducibility, examined across an expanded panel of 91 isolates, demonstrated >98% PCR positivity rate with minimal non-specific amplification.
- LOH was observed in TcV and TcVI for two targets (*Met-II* and *TcMPX*), resulting in phylogenetic incongruence. However, the latter target was included in both final panels to distinguish between the two hybrid lineages.
- This formalized MLST scheme represents a highly discriminatory strain typing technique and new ‘gold standard’ for routine *T. cruzi* lineage assignment (four loci) and higher resolution diversity studies (seven loci).
- Future work will include application of this MLST scheme to larger field cohorts for more comprehensive population genetics studies.
Please be aware that one cover sheet must be completed for each ‘Research Paper’ included in a thesis.

1. For a ‘research paper’ already published

   1.1. Where was the work published?  PLoS Neglected Tropical Diseases

   1.2. When was the work published?  2014

   1.2.1. If the work was published prior to registration for your research degree, give a brief rationale for its inclusion
   ..........................................................................................................................
   ..........................................................................................................................

   1.3. Was the work subject to academic peer review?  Yes

   1.4. Have you retained the copyright for the work? Yes

   If yes, please attach evidence of retention.

   If no, or if the work is being included in its published format, please attach evidence of permission from copyright holder (publisher or other author) to include work

2. For a ‘research paper’ prepared for publication but not yet published

   2.1. Where is the work intended to be published?  .................................................................

   2.2. Please list the paper’s authors in the intended authorship order
   ..............................................................................................................................................

   2.3. Stage of publication – Not yet submitted / Submitted / Undergoing revision from peer reviewers’ comments / In press

3. For multi-authored work, give full details of your role in the research included in the paper and in the preparation of the paper. (Attach a further sheet if necessary)

   The candidate generated complete sequence data for five of thirteen housekeeping genes under evaluation (Met-II, Met-III, RB19, TcAPX and TcMPX) and participated in data interpretation, analysis and initial manuscript drafting.

NAME IN FULL (Block Capitals)  LOUISA ALEXANDRA MESSENGER

STUDENT ID NO:  223021

CANDIDATE’S SIGNATURE  

Date 10/2/15

SUPERVISOR/SENIOR AUTHOR’S SIGNATURE (3 above)  

Improving health worldwide  www.lshtm.ac.uk
Introduction

Trypanosoma cruzi, the protozoan causative agent of Chagas disease, is a monophyletic and genetically heterogeneous taxon, with at least six phylogenetic lineages formally recognised as Discrete Typing Units (DTUs), TcI–TcVI [1], or near-clades (clades that are blurred by infrequent inter-lineage genetic recombination, [2]). T. cruzi is considered to have a predominantly clonal population structure but with at least some intra-lineage recombination [3,4,5,6]. TcI and TcII are the most genetically distant groups, and the evolutionary origins of TcIII and TcIV remain controversial. Based on sequencing of individual nuclear genes Westenberger et al. [7] suggested that an ancient hybridisation event occurred between TcI and TcII followed by a long period of clonal propagation leading to the extant TcIII and TcIV. Alternatively, de Freitas et al. [8] suggested that TcIII and TcIV have a separate evolutionary ancestry with mitochondrial sequences that are similar to each other but distinct from both TcI and TcII. Recently, Flores-Lopez and Machado [9] proposed that TcII and TcIV have no hybrid origin. Based on the sequence of 32 genes, they strongly suggested that TcI, TcIII and TcIV are clustered into a major clade that diverged from TcII around 1–2 millions of years ago. Less controversially, it is clear that TcV and TcVI, both overwhelmingly represented in the domestic transmission cycles in the Southern Cone region of South America, are hybrid lineages sharing haplotypes from both TcII and TcIII, with both DTUs retaining the mitochondrial genome of TcIII [8,10]. Recent phylogenetic studies suggest that the emergence of the hybrid lineages TcV and TcVI may have occurred within the last 60,000 years [11]. Reliable DTU identification and the potential for high resolution investigation of genotypes at the intra DTU level are of great interest for epidemiological, host association, clinical and phylogenetic studies. Historically, a plethora of typing techniques have been applied to T. cruzi. Initial pioneering work applied multilocus enzyme electrophoresis (MLEE) techniques [12,13,14,15,16,17,18,19,20] revealing the remarkable genetic diversity of T. cruzi. However, the advent of high-throughput sequencing in recent years has revolutionised molecular epidemiology. Whole genome sequencing or long-read metagenomic shotgun sequencing [16,21,22] has been used to investigate the epidemiology of T. cruzi [17,23]. The aetiological agent of Chagas disease possess extensive genetic diversity. This has led to the development of a plethora of molecular typing methods for the identification of both the known major genetic lineages and for more fine scale characterisation of different multilocus genotypes within these major lineages. Whole genome sequencing applied to large sample sizes is not currently viable and multilocus enzyme electrophoresis, the previous gold standard for T. cruzi typing, is laborious and time consuming. In the present work, we present an optimized Multilocus Sequence Typing (MLST) scheme, based on the combined analysis of two recently proposed MLST approaches. Here, thirteen concatenated gene fragments were applied to a panel of T. cruzi reference strains encompassing all known genetic lineages. Concatenation of 13 fragments allowed assignment of all strains to the predicted Discrete Typing Units (DTUs), or near-clades, with the exception of one strain that was an outlier for TcV, due to apparent loss of heterozygosity in one fragment. Monophyly for all DTUs, along with robust bootstrap support, was restored when this fragment was subsequently excluded from the analysis. All possible combinations of loci were assessed against predefined criteria with the objective of selecting the most appropriate combination of between two and twelve fragments, for an optimized MLST scheme. The optimum combination consisted of 7 loci and discriminated 21 out of 25 genotypes. We propose that the seven-fragment MLST scheme could be used as a gold standard for T. cruzi typing, against which other typing approaches, particularly single locus approaches or systematic PCR assays based on amplicon size, could be compared.
Parasite strains and DNA isolation

Methods

Parasite strains and DNA isolation

Twenty five cloned reference strains belonging to the six known DTUs were examined (Table 1). These strains have been widely used as reference strains in many previous studies, and are regularly examined in our laboratory by Multilocus Enzyme Electrophoresis (MLEE). Parasite stocks were cultivated at 28°C in liver infusion tryptose (LIT) supplemented with 1% hemin, 10% fetal bovine serum, 100 units/ml of penicillin, and 100 µg/mL of streptomycin or in supplemented RPMI liquid medium.

MLST loci

Initially a total of 19 gene fragments were considered, 10 housekeeping genes previously described by Lauthier et al. [50] [Glutathione peroxidase (GPX), 3-Hydroxi-3-methylglutaryl-CoA reductase (HMGCoAR), Piruvate dehydrogenase component E1 subunit alfa (PDH), Small GTP-binding protein Rab7 (GTP), Serine/threonine-protein phosphatase PPI (STPP2), Rho-like GTP binding protein (RHO1), Glucose-6-phosphate isomerase (GPI), Superoxide dismutase A (SODA), Superoxide dismutase B (SODB) and Leucine aminopeptidase (LAP)]; and 9 gene fragments from Yeo et al. [51] [ascorbate-dependent haemoperoxidase (TcAPX), dihydrofolate reductase-thymidylate synthase (DHFR-TS), glutathione-dependent peroxidase II (TcGPXII), mitochondrial peroxi- idase (TcMPX), trypanothione reductase (TR), RNA-binding protein-19 (RB19), metacyclin-II (Met-II), metacyclin-III (Met-III) and LYT1]. However, 6 of them were discarded based on initial findings [50,51]. Although some of the excluded targets were informative, they were not amenable for routine use. More specifically, LYT1 was discarded due to unreliable PCR amplification and sequencing despite multiple attempts at optimization; TR, DHFR-TS and TcAPX were also deemed unsuitable as internal sequencing primers were required; finally, Met-III and TcGPXII were also excluded because generated non-specific PCR products with some isolates.

The final 13 gene fragments assessed included 3 fragments described by Yeo et al. [51] and the 10 housekeeping genes previously described by Lauthier et al. [50]. These were: TcMPX, RB19, Met-II, SODA, SODB, LAP, GPI, GPX, PDH, HMGCoAR, RHO1, GTP and STPP2. For the 13 loci under study, searches in the CL-Brener and Sylvio X10 genomes (http://tritrypdb.org/tritrypdb/), using the primer sequences as well as the fragment sequences as query, displayed single matches in all cases. Chromosome location, primer sequences and ampiclon size for each target are shown in Table 2. Nucleotide sequences for all the analysed MLST targets are available from GenBank under the following accession numbers: JN129501-JN129502, JN129511-JN129518, JN129523-JN129524, JN129534-JN129535, JN129544-JN129551, JN129556-JN129557, JN129567-JN129568, JN129577-JN129584, JN129598-JN129600, JN129601-JN129610, JN129616-JN129617, JN129622-JN129623, JN129633-JN129634, JN129643-JN129650, JN129653-JN129654, JN129666-JN129667, JN129667-JN129683, JN129688-JN129689, JN129689-JN129690, JN129690-JN129697, JN129697-JN129698, JN129698-JN129699, JN129699-JN129700, JN129700-JN129709, JN129710-JN129716, JN129721-JN129722, JN129732-JN129733, JN129748-JN129749, JN129749-JN129755, JN129765-JN129766, JN129767-JN129782, JN129787-JN129788, JN129798-JN129799, JN129808-JN129815, JN129820-JN129821, KF889442-KF898646. Additionally, we used T. cruzi marinkellei as outgroup. Sequence data of the selected targets for T. cruzi marinkellei were obtained from TriTrypDB (http://tritrypdb.org), under the following accession IDs: TeMARK_CONTIG_2666, TeMARK_CONTIG_670, TeMARK_CONTIG_1104, TeMARK_2068, TeMARK_3409, TeMARK_5695, TeMARK_9874, TeMARK_4984, TeMARK_5926, TeMARK_8923, TeMARK_CONTIG_1818 and TeMARK_2666.

Molecular methods

PCRs were performed in 50 µl reaction volumes containing 100 ng of DNA, 0.2 µl of each primer, 1 U of goTaq DNA polymerase (Promega), 10 µl of buffer (supplied with the goTaq
polymerase) and a 50 μM concentration of each deoxynucleoside triphosphate (Promega). Amplification conditions for all targets were: 5 min at 94°C followed by 35 cycles of 94°C for 1 min; 55°C 1 min, and 72°C for 1 min, with a final extension at 72°C for 5 min. Amplified fragments were purified (QiAquick, Qiagen) and sequenced in both directions (ABI PRISM 310 Genetic Analyzer or ABI PRISM 377 DNA Sequencers, Applied Biosystems) using standard protocols. Primers used for sequencing were identical to those used in PCR amplifications. In order to assess reproducibility, each PCR amplification was performed twice.

### Data analysis

MLST data were analysed with MLSTest software [http://ipe.unsa.edu.ar/software] [52] with the objective of identifying the most resolutive and minimum number of targets for unequivocal DTU assignment and potential fine scale characterisation. MLSTest contains a suite of MLST data specific analytical tools. Briefly, single nucleotide polymorphisms (SNPs) were identified in all loci in MLSTest alignment viewer. Typing efficiency (TE) was calculated using the same software. TE for a determined locus is calculated as the number of identified genotypes divided by the number of polymorphic sites in this locus. Additionally, discriminatory power, defined as the probability that two strains are distinguished when chosen at random from a population of unrelated strains [53], was determined for each target (Table 3).

Sequence data were concatenated and Neighbour Joining phylogenetic trees were generated by using uncorrected p-distances. Heterozygous sites were handled in the analyses using two different methods. First, a SNP duplication method described by Yeo et al. and Tavanti et al. [51,54] was implemented. Briefly, the SNP duplication method involves the elimination of monomorphic sites and duplication of polymorphisms in order to “resolve” the heterozygous sites. As an example, a homozygous variable locus scored as C (cytosine) will be modified by CC; while a heterozygous locus, for example Y (C or T, in accordance with IUPAC nomenclature), will be scored as CT. Alternatively, heterozygous SNPs were considered as average states. In more detail, the genetic distance between T and Y (heterozygosity composed of T and C) is considered as the mean distance between the T and the possible resolutions of Y (distance T-T = 0 and distance T-C = 1, average distance = 0.5, see [53] and MLSTest 1.0 manual at http://www.ipe.unsa.edu.ar/software for further details). Statistical support was evaluated by 1000 bootstrap replications. Overall phylogenetic incongruence among loci (by comparison with the concatenated topology) was assessed by the Incongruence Length Difference Test using the BIO-Neighbour Joining method [BIONJ-ILD, [55]] and evaluated by a permutation test with 1,000 replications. Briefly, the ILD evaluates whether the observed incongruence among fragments is higher than that expected by random unstructured homoplasy across the different fragments. A statistical significant ILD p value indicates that many sites, in at least one fragment, support a phylogeny that

### Table 1. Cohort of clonal reference isolates representing the six known T. cruzi lineages (DTUs).

<table>
<thead>
<tr>
<th>Strain</th>
<th>DTU</th>
<th>Origin</th>
<th>Host</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. X10c1</td>
<td>Tcl</td>
<td>Belém, Brazil</td>
<td>Homo sapiens</td>
</tr>
<tr>
<td>2. Cutia c1</td>
<td>Tcl</td>
<td>Esprírito Santo, Brazil</td>
<td>Dasypodinae aguti</td>
</tr>
<tr>
<td>3. Sp104 c1</td>
<td>Tcl</td>
<td>Region IV, Chile</td>
<td>Triatoma spinoff</td>
</tr>
<tr>
<td>4. P209 c193</td>
<td>Tcl</td>
<td>Sucre, Bolivia</td>
<td>Homo sapiens</td>
</tr>
<tr>
<td>5. OPS21 c11</td>
<td>Tcl</td>
<td>Cojedes, Venezuela</td>
<td>Homo sapiens</td>
</tr>
<tr>
<td>6. 92101601P c1</td>
<td>Tcl</td>
<td>Georgia, USA</td>
<td>Didelphis marsupialis</td>
</tr>
<tr>
<td>7. TU18 c193</td>
<td>TcII</td>
<td>Potosi, Bolivia</td>
<td>Triatoma infestans</td>
</tr>
<tr>
<td>8. CBB c3</td>
<td>TcII</td>
<td>Region IV, Chile</td>
<td>Homo sapiens</td>
</tr>
<tr>
<td>9. Mas c1</td>
<td>TcII</td>
<td>Federal District, Brazil</td>
<td>Homo sapiens</td>
</tr>
<tr>
<td>10. IVV c4</td>
<td>TcII</td>
<td>Region IV, Chile</td>
<td>Homo sapiens</td>
</tr>
<tr>
<td>11. Esm c3</td>
<td>TcII</td>
<td>São Felipe, Brazil</td>
<td>Homo sapiens</td>
</tr>
<tr>
<td>12. M6241 c6</td>
<td>TcIII</td>
<td>Belém, Brazil</td>
<td>Homo sapiens</td>
</tr>
<tr>
<td>13. CM17</td>
<td>TcIII</td>
<td>Meta, Colombia</td>
<td>Dasypodinae sp.</td>
</tr>
<tr>
<td>15. X109/2</td>
<td>TcIII</td>
<td>Miktlawaya, Paraguay</td>
<td>Canis familiaris</td>
</tr>
<tr>
<td>16. 92122102R</td>
<td>TcIV</td>
<td>Georgia, USA</td>
<td>Procyon lotor</td>
</tr>
<tr>
<td>17. Canilli c1</td>
<td>TcIV</td>
<td>Belém, Brazil</td>
<td>Homo sapiens</td>
</tr>
<tr>
<td>18. Dog Theis</td>
<td>TcIV</td>
<td>USA</td>
<td>Canis familiaris</td>
</tr>
<tr>
<td>19. Mn c2</td>
<td>TcV</td>
<td>Region IV, Chile</td>
<td>Homo Sapiens</td>
</tr>
<tr>
<td>20. Bug 2148 c1</td>
<td>TcV</td>
<td>Rio Grande do sul, Brazil</td>
<td>Triatoma infestans</td>
</tr>
<tr>
<td>21. SO3 c5</td>
<td>TcV</td>
<td>Potosi, Bolivia</td>
<td>Triatoma infestans</td>
</tr>
<tr>
<td>22. SC45 c1</td>
<td>TcV</td>
<td>Santa-Cruz, Bolivia</td>
<td>Triatoma infestans</td>
</tr>
<tr>
<td>23. CL Brener</td>
<td>TcVI</td>
<td>Rio Grande do sul, Brazil</td>
<td>Triatoma infestans</td>
</tr>
<tr>
<td>24. P63 c1</td>
<td>TcVI</td>
<td>Miktlawaya, Paraguay</td>
<td>Triatoma infestans</td>
</tr>
<tr>
<td>25. Tula c2</td>
<td>TcVI</td>
<td>Talahuem, Chile</td>
<td>Homo sapiens</td>
</tr>
</tbody>
</table>
is contradicted by other fragments. In order to localize significant incongruent branches in concatenated data we used the Neighbour Joining based Localized Incongruence Length Difference (NJ-LILD) test available in MLSTest. NJ-LILD is a variant of the ILD test that allows localizing incongruence at branch level. All combinations from 2 to 12 fragments were analysed using the scheme optimisation algorithm in MLSTest which identifies the combination of loci producing the maximum number of diploid sequence types (DSTs). Three main sequential criteria were applied to select the optimum combination of loci: firstly, monophyly of DTUs and lineage assignment; secondly, robust bootstrap values for the six major DTUs (1000 replications); and thirdly detection of genetic diversity at the intra-DTU level.

Results

PCR amplification and sequencing
All 13 gene fragments were successfully amplified using identical PCR reaction conditions (see methods) which generated discrete PCR fragments. PCR amplifications of the 13 targets were applied to an extended panel of 90 isolates obtaining more than 98% of positive PCR and amplifications produced strong amplicons and an absence of non-specific products (data not shown). We obtained amplicons of the expected length for all the assayed targets and for all the examined strains. Amplification for various DNA template concentrations was assayed via serial dilution. No difference in PCR amplifications were obtained when DNA concentrations from 20 to 100 ng were used. A total of 5,121 bp of sequence data were analysed for each strain (Table 2). There were no gaps in sequences. The number of polymorphic sites (Table 3) for each of the different fragments varied from 8 (STPP2) to 40 (Met-II). STTP2 showed the lowest discriminatory power (describing just 5 different genotypes in the dataset). Rb19 was the fragment with the highest discriminatory power identifying 21 distinct genotypes in the dataset.

Optimized scheme for MLST
Initially, Neighbor Joining trees were generated from concatenated sequences across the 13 prescreened loci which identified four monophyletic DTUs with robust bootstrap support (TcI, TcII, TcIII, TcIV).
TcII, TcIII, TcIV, bootstrap >98%). TcVI was also monophyletic but with a relatively low support (Figure 1). Additionally, TcV was paraphyletic with MncI2 as an outlier. The concatenated 13 fragments differentiated all 25 reference strains in terms of DSTs. We observed that bootstrap values were slightly different between the two methods (SNP duplication and average alignments) as they manage heterozygous sites differently. Values were higher for the SNP duplication method in most branches (Figure 1, branch values highlighted in blue) as a consequence of base duplication, which modifies the alignment and increases the informative sites used for bootstrapping. To avoid the potential for methodologically elevated bootstraps, the average states method was implemented for further analyses. From the selected 13 loci, all possible combinations of 2 to 12 loci were analysed (8,177 combinations) required for DTU assignment while maintaining DST identification. All combinations of 3 and 4 fragments (1,001 combinations) from the panel of 13 fragments were analysed as described above. A reduced MLST panel incorporating TcMPX, HMCOAR, RH01 and GPl (four loci) produced the highest bootstrap values for DTU assignment across the DTUs, TcI (99.9), TcII (100), TcIII (99.3), TcIV (86.7), TcV (100) and TcVI (96.8) (Figure 3), and discriminated 19 of 25 DSTs. Other combinations showed higher discriminatory power but presented with lower bootstrap values (data not shown). The TcMPX locus exhibits an apparent loss of heterozygosity (LOH) in the hybrid DTU TcV, retaining the TcII-like allele but not the TcIII allele. Therefore DTU assignment using TcMPX alone would not assign a TcV isolate correctly. However concatenation of TcMPX with HMCOAR, RH01 and GPI allow distinguishing TcV from TcII.

### Reduced scheme for DTU assignment

Attempts were made to reduce the number of fragments required for DTU assignment while maintaining DST identification. All combinations of 3 and 4 fragments (1,001 combinations) from the panel of 13 fragments were analysed as described above. A reduced MLST panel incorporating TcMPX, HMCOAR, RH01 and GPl (four loci) produced the highest bootstrap values for DTU assignment across the DTUs, TcI (99.9), TcII (100), TcIII (99.3), TcIV (86.7), TcV (100) and TcVI (96.8) (Figure 3), and discriminated 19 of 25 DSTs. Other combinations showed higher discriminatory power but presented with lower bootstrap values (data not shown). The TcMPX locus exhibits an apparent loss of heterozygosity (LOH) in the hybrid DTU TcV, retaining the TcII-like allele but not the TcIII allele. Therefore DTU assignment using TcMPX alone would not assign a TcV isolate correctly. However concatenation of TcMPX with HMCOAR, RH01 and GPI allow distinguishing TcV from TcII.

### Inter and intra DTU phylogenies

Topologies obtained for the 7 and 4 loci combinations (Figures 2 and 3, respectively) were similar to the 13 loci scheme, showing consistently the two major groups (TcI-TcII-TcIV and TcII-TcV-TcVI) supported by high bootstrap values, even when trees were rooted using TcMB7 (Figure 1). The primary difference between the 13 target concatenated phylogenies and the trees obtained for the 7 and 4 targets was that for the 13 concatenated targets TcV was paraphyletic, showing the MncI2 strain as an outlier. Regarding inter-DTU relationships, the analysis of the concatenated 13 fragments divided DTUs into two major clusters: one composed by TcI, TcIII and TcIV, with a bootstrap value of 100%; while the remaining group containing TcII, TcV and TcVI was supported by lower bootstrap values (<70%), possibly due to presence of the two hybrid DTUs (TcV and TcVI) (Figure 1). Within clusters, internal topologies were supported with relatively high but variable bootstrap values with 4, 7 and 13 loci combinations and generally consistent intralineage topologies (Figures 1, Figure 2, Figure 3), although the panel of 25 reference

### Table 3. T. cruzi MLST targets.

<table>
<thead>
<tr>
<th>Gene fragment</th>
<th>No. of genotypes</th>
<th>No. of polymorphic sites</th>
<th>Typing efficiency</th>
<th>DP</th>
</tr>
</thead>
<tbody>
<tr>
<td>GPI</td>
<td>9</td>
<td>18</td>
<td>0.500</td>
<td>0.889</td>
</tr>
<tr>
<td>HMCOAR</td>
<td>15</td>
<td>20</td>
<td>0.750</td>
<td>0.954</td>
</tr>
<tr>
<td>RH01</td>
<td>13</td>
<td>23</td>
<td>0.565</td>
<td>0.914</td>
</tr>
<tr>
<td>TcMPX</td>
<td>11</td>
<td>12</td>
<td>0.917</td>
<td>0.905</td>
</tr>
<tr>
<td>LAP</td>
<td>13</td>
<td>16</td>
<td>0.812</td>
<td>0.942</td>
</tr>
<tr>
<td>SODB</td>
<td>12</td>
<td>9</td>
<td>1.333</td>
<td>0.914</td>
</tr>
<tr>
<td>RB19</td>
<td>21</td>
<td>26</td>
<td>0.808</td>
<td>0.985</td>
</tr>
<tr>
<td>GPX</td>
<td>12</td>
<td>16</td>
<td>0.750</td>
<td>0.908</td>
</tr>
<tr>
<td>PDH</td>
<td>11</td>
<td>15</td>
<td>0.733</td>
<td>0.920</td>
</tr>
<tr>
<td>GTP</td>
<td>10</td>
<td>18</td>
<td>0.556</td>
<td>0.905</td>
</tr>
<tr>
<td>SODA</td>
<td>10</td>
<td>10</td>
<td>1.000</td>
<td>0.880</td>
</tr>
<tr>
<td>STPP2</td>
<td>5</td>
<td>8</td>
<td>0.625</td>
<td>0.585</td>
</tr>
<tr>
<td>Met-II</td>
<td>19</td>
<td>40</td>
<td>0.475</td>
<td>0.978</td>
</tr>
</tbody>
</table>

DP: Discriminatory Power according to [53].
1 Number of genotypes per polymorphic site,
* included in the seven loci scheme,
* included in the four loci scheme.

*doi:10.1371/journal.pntd.0003117.t003*
strains would need to be expanded further for assessment of fine scale intralineage associations.

**Discussion**

Thirteen gene fragments were assessed in an optimised MLST scheme which is a combination of targets from two recently separately proposed schemes [50,51]. Here we evaluated the optimal combination of loci based on three main sequential criteria: first, assignment to the expected DTU; second, to attain robust bootstrap values for the six major DTUs, and third to detect intra-DTU diversity. For the first time we propose an optimised MLST scheme, validated against a panel representing all known lineages, for characterisation of *Trypanosoma cruzi* isolates. However, it should be emphasized that this MLST scheme is proposed as a typing method for *T. cruzi* isolates but not as a typing method to be used directly on biological samples as blood, tissues or Triatomine feces, for which more sensitive and simpler methods are needed. Moreover, we have performed assays with the purpose of determining the limit of detection of each gene fragment on blood and triatomines feces (data not shown) and we found that none of these targets are suitable for detecting *T. cruzi* in the normal concentration found in natural biological samples.

As a result of our data analyses, we obtained one combination of 7 loci and one combination of only 4 targets which most closely adhered to the selection criteria described above. It is worth noting that the three used criteria for selecting optimum combination of targets are sequential; it means that there is a hierarchical order of these criteria. In first place, we look for obtaining monophyly for the six DTUs and accurate lineage assignment of each examined strain. In a second place, we look for obtaining robust bootstrap values for each of the six major DTUs. Finally, we expect detecting genetic diversity at the intra-DTU level. In this context, due to the hierarchical order of the criteria of selection of loci, the selected combinations will optimise the number of DSTs but subordinated to the two previous criteria. Theoretically, using these criteria, we could obtain a combination of loci that does not give the maximum number of DST for a determined DTU, because our algorithm previously prioritized obtaining monophyly and strong bootstrap values for the six DTUs. This was the case for the selected 4-loci scheme (which differentiated 19 from 25 strains). In spite of this, the selected 7-loci combination that we propose, allow us to differentiate the 25 examined strains, i.e. the maximum possible number of DSTs. The results illustrate that MLST is a highly discriminatory strain-typing technique. From these data we suggest that the 7 locus scheme provides scope for both lineage assignment and diversity studies, generating robust bootstrap values for distance based phylogenies and that a reduced panel of only four targets is sufficient for assignment to DTU level. For population genetics scale analyses and detailed epidemiological...
The phylogenetic associations among DTUs TcI, TcII, TcIII and TcIV are debatable. Split affinities and incongruence have been observed in nuclear phylogenies [7,8,51,56]. One interpretation of phylogenetic incongruence is genetic recombination, although due to the highly plastic nature of the T. cruzi genome other causes are also possible. Mutation rates and gene conversion may create distinct levels of sequence diversity [57]. Here, concatenated phylogenies showed a partition into two main clusters for all gene combinations tested, the first consisting of TcI, TcIII and TcIV (bootstrap value = 100%) and the second composed of TcII, TcV and TcVI (bootstrap value <70%). The presence of the two known hybrid lineages (TcV and TcVI) generated artifactual phylogenetic structuring and excluding these representatives revealed clustering of DTUs TcI, TcII, TcIII and TcIV, indicating that TcI is a closer affinity to TcIII than to TcIV. TcII is the most genetically distant group which is in agreement with previous findings [9,10,51]. In addition, it would be interesting to analyze in the future the new lineage described as TcBat [58] using the MLST scheme proposed here, since it could shed light on the phylogenetical position of this interesting lineage.

LOH observed in Met-II and TcMPX gene fragments affecting the hybrid lineages TcV and TcVI has potentially significant consequences for MLST and lineage assignment [51]. Isolates affected retain the TcII like allele and would be misassigned in single locus characterisation. For example, hybrid isolates TcV would be assigned to TcII based on TcMPX sequencing due to apparent LOH. Despite this LOH the TcVI isolate would be assigned to TcII based on single locus characterisation. For example, hybrid isolates TcV and TcVI has potentially significant consequences for MLST and lineage assignment [51]. Isolates affected retain the TcII like allele and would be misassigned in single locus characterisation. For example, hybrid isolates TcV would be assigned to TcII based on TcMPX sequencing due to apparent LOH. Despite this LOH the TcVI isolate would be assigned to TcII based on single locus characterisation.
minimal cleanup and are suitable for sequencing. Although in the current protocol, we recommend purifying PCR products with a suitable commercial kit (Qiagen), in most cases, this was not necessary and sequencing was performed directly from the PCR product. The exception was TcGPIX, and very occasionally SODA produced nonspecific products, neither of which are included in final recommended panels. Although the two previously published MLST [50,51] schemes showed promise in identifying diversity, some of the gene targets were not amenable for routine use. For example, LYT1 was discarded due to unreliable amplification and DHFR-TS due to the need for internal primers. Therefore further optimisation performed here was necessary for practical use. An important criterion for choosing targets was identifying those that used the same primers for both PCR amplification and sequencing to maintain simplicity and reduce costs.

Taken together, we propose a MLST scheme validated against a panel representing all of the known lineages of T. cruzi. We propose that a 7 loci MLST scheme could provide the basis for robust DTU assignment and strain diversity studies of new isolates and a reduced 4 loci scheme for lineage assignment. Importantly, the sequence data generated can be utilised for a wide range of downstream analyses, including the resolution of haplotypes for recombination analysis, population genetics analyses, and other statistical approaches to the phyloepidemiological study of T. cruzi.

Finally, we propose that the seven-fragment MLST scheme could be used as a gold standard for T. cruzi typing, against which other typing approaches, particularly single locus approaches or systematic PCR assays based on ampiclon size, could be compared.

Acknowledgments

We are grateful to Alejandro D. Uncos, Federico Ramos and Maria Celia Mora for their technical support.

Author Contributions

Conceived and designed the experiments: PD MY NT JYL MAM LAM MT CB MSL MDL. Performed the experiments: PD MY JYL LAM NT MMR PGR AMAD CPB MDL. Analyzed the data: NT PD MY JYL LAM. Contributed reagents/materials/analysis tools: PD MAM. Wrote the paper: PD MY NT. Designed the software used in analysis: NT.

References


3.3.3 Development and evaluation of mitochondrial MLST targets

To evaluate the suitability of multi-copy maxicircle genes as phylogenetic markers, the level of intra-clone maxicircle genetic diversity was examined in silico using Illumina sequencing reads, generated as part of the Sylvio X10/1 genome project (Franzén et al., 2011). Sequencing reads were assembled and mapped onto the Sylvio X10/1 maxicircle genome to identify the presence of any minor maxicircle haplotypes (evidence of mitochondrial heteroplasmy), which could represent a potential source of genotyping error.

Subsequently, to describe intra-DTU diversity, a high resolution mtMLST scheme was developed based on ten mitochondrial gene fragments. This mtMLST scheme was evaluated using a panel of 32 TcI isolates, which was representative of the endemic geographical range of this DTU. In parallel, isolates were also characterized using nuclear loci and the extent of nuclear-mitochondrial phylogenetic incongruence was examined to detect incidences of genetic exchange.

This study is reported in full below by Messenger et al., 2012.

In summary:

- Illumina sequencing data from the TcI genome strain revealed multiple minor heteroplasmic maxicircles (~10 fold lower abundance compared to the consensus genome) within an individual parasite that were, however, not sufficiently divergent to represent a major source of genotyping error.
- Robust amplification of the ten maxicircle gene fragments was confirmed across all six T. cruzi DTUs.
- The resolutive power of the mtMLST scheme was equivalent or superior to nuclear markers routinely used to describe intra-TcI diversity (GPI, SL-IR and MLMT).
- Comparison of nuclear and mitochondrial topologies revealed multiple mitochondrial introgression events between major lineages in Venezuela and North America (TcI and TcII/TcIV) and within TcI populations in Argentina, Bolivia and Brazil.
- Absence of reciprocal nuclear hybridization suggests that mitochondrial introgression may occur independently of nuclear recombination. These observations also highlight the importance of using mitochondrial markers to identify cryptic diversity and recombination events which were undetectable using conventional nuclear loci.
- Gross phylogenetic incongruence indicates that genetic exchange is contemporary and geographically widespread among natural TcI populations, a conclusion which challenges the traditional paradigm of clonality in T. cruzi.
- The mtMLST scheme provided a powerful approach to genotyping at the sub-lineage level. The combined nuclear-mitochondrial strategy will facilitate attempts to address epidemiologically important hypotheses in conjunction with intensive spatio-temporal parasite sampling.
COVER SHEET FOR EACH ‘RESEARCH PAPER’ INCLUDED IN A RESEARCH THESIS

Please be aware that one cover sheet must be completed for each ‘Research Paper’ included in a thesis.

1. For a ‘research paper’ already published

1.1. Where was the work published? PLoS Neglected Tropical Diseases

1.2. When was the work published? 2012

1.2.1. If the work was published prior to registration for your research degree, give a brief rationale for its inclusion

1.3. Was the work subject to academic peer review? Yes

1.4. Have you retained the copyright for the work? Yes/No

   If yes, please attach evidence of retention.

   If no, or if the work is being included in its published format, please attach evidence of permission from copyright holder (publisher or other author) to include work

2. For a ‘research paper’ prepared for publication but not yet published

2.1. Where is the work intended to be published?

2.2. Please list the paper’s authors in the intended authorship order

2.3. Stage of publication – Not yet submitted / Submitted / Undergoing revision from peer reviewers’ comments / In press

3. For multi-authored work, give full details of your role in the research included in the paper and in the preparation of the paper. (Attach a further sheet if necessary)

   The candidate designed the study and performed all mitochondrial sequencing including assembly of the Sylvio X10/1 mitochondrial genome with guidance from collaborators at the Karolinska Institutet, Sweden. The candidate also generated the microsatellite data; analyzed all data; drafted the manuscript and was responsible for final manuscript revisions for publication.

NAME IN FULL (Block Capitals) LOUISA ALEXANDRA MESSENGER

STUDENT ID NO: 223021

CANDIDATE’S SIGNATURE: [Signature] Date 10/2/15

SUPERVISOR/SENIOR AUTHOR’S SIGNATURE (3 above)
Multiple Mitochondrial Introgression Events and Heteroplasmy in *Trypanosoma cruzi* Revealed by Maxicircle MLST and Next Generation Sequencing

Louisa A. Messenger1*, Martin S. Llewellyn1, Tapan Bhattacharyya1, Oscar Franžén2, Michael D. Lewis1, Juan David Ramírez3, Hernan J. Carrasco4, Björn Andersson2, Michael A. Miles1

1Department of Pathogen Molecular Biology, Faculty of Infectious and Tropical Diseases, London School of Hygiene and Tropical Medicine, London, United Kingdom, 2Science for Life Laboratory, Department of Cell and Molecular Biology, Karolinska Institutet, Stockholm, Sweden, 3Centro de Investigaciones en Microbiología y Parasitología Tropical, CIMPAT, Universidad de los Andes, Bogotá, Colombia, 4Instituto de Medicina Tropical, Facultad de Medicina, Universidad Central de Venezuela, Los Chaguaramos, Caracas, Venezuela

Abstract

**Background:** Mitochondrial DNA is a valuable taxonomic marker due to its relatively fast rate of evolution. In *Trypanosoma cruzi*, the causative agent of Chagas disease, the mitochondrial genome has a unique structural organization consisting of 20–50 maxicircles (~20 kb) and thousands of minicircles (0.5–10 kb). *T. cruzi* is an early diverging protist displaying remarkable genetic heterogeneity and is recognized as a complex of six discrete typing units (DTUs). The majority of infected humans are asymptomatic for life while 30–35% develop potentially fatal cardiac and/or digestive syndromes. However, the relationship between specific clinical outcomes and *T. cruzi* genotype remains elusive. The availability of whole genome sequences has driven advances in high resolution genotyping techniques and re-invigorated interest in exploring the diversity present within the various DTUs.

**Methodology/Principal Findings:** To describe intra-DTU diversity, we developed a highly resolutive maxicircle multilocus sequence typing (mtMLST) scheme based on ten gene fragments. A panel of 32 TcI isolates was genotyped using the mtMLST scheme, GPI, mini-exon and 25 microsatellite loci. Comparison of nuclear and mitochondrial data revealed clearly incongruent phylogenetic histories among different geographical populations as well as major DTUs. In parallel, we exploited read depth data, generated by Illumina sequencing of the maxicircle genome from the TcI reference strain Sylvio X10/1, to provide the first evidence of mitochondrial heteroplasmy (heterogeneous mitochondrial genomes in an individual cell) in *T. cruzi*.

**Conclusions/Significance:** mtMLST provides a powerful approach to genotyping at the sub-DTU level. This strategy will facilitate attempts to resolve phenotypic variation in *T. cruzi* and to address epidemiologically important hypotheses in conjunction with intensive spatio-temporal sampling. The observations of both general and specific incidences of nuclear-mitochondrial phylogenetic incongruence indicate that genetic recombination is geographically widespread and continues to influence the natural population structure of TcI, which challenges the traditional paradigm of clonality in *T. cruzi*.


Editor: Yara M. Traub-Cseko, Instituto Oswaldo Cruz, Fiocruz, Brazil

Received October 4, 2011; Accepted February 15, 2012; Published April 10, 2012

Copyright: © 2012 Messenger et al. This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Funding: This work was supported by the Wellcome Trust and the European Commission Framework Programme Project “Comparative epidemiology of genetic lineages of *Trypanosoma cruzi*” ChagasEpiNet, Contract No. 223034. LAM is supported by the BBSRC. The human *T. cruzi* Venezuelan isolates were obtained with support from Grant FONACIT G-2005000827. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: louisa.messenger@lshtm.ac.uk

Introduction

Mitochondrial genes are among the most popular markers for the reconstruction of evolutionary ancestries and resolution of phylogeographic relationships [1]. Their pervasive use in population genetics can be attributed to several intrinsic characteristics, notably, their high copy number, small size (~15–20 kb) and faster mutation rate (compared with nuclear DNA). In addition, their widespread application is founded on the assumptions that mitochondrial genomes are homoplasmic, uniparentally inherited and lack homologous recombination [2]. However, with technological advances affording increased sensitivity and greater sample throughput, a growing number of reports of heteroplasm (heterogeneous mitochondrial genomes in an individual cell), introgression and inter-molecular recombination are challenging what was previously regarded as a strict set of rules for eukaryotic mitochondrial inheritance.

Chagas disease remains the most important parasitic infection in Latin America, where an estimated 10–12 million individuals are infected, with a further 80 million at risk [3]. The aetiological agent,
Author Summary

Chagas disease, caused by the protozoan parasite Trypanosoma cruzi, is an important public health problem in Latin America. While molecular techniques can differentiate the major T. cruzi genetic lineages, few have sufficient resolution to describe diversity among closely related strains. The online availability of three mitochondrial genomes allowed us to design a multilocus sequence typing (mtMLST) scheme to exploit these rapidly evolving markers. We compared mtMLST with current nuclear typing tools using isolates belonging to the oldest and most widely occurring lineage Tcl. T. cruzi is generally believed to reproduce clonally. However, in this study, distinct branching patterns between mitochondrial and nuclear phylogenetic trees revealed multiple incidences of genetic exchange within different geographical populations and major lineages. We also examined Illumina sequencing data from the Tcl genome strain which revealed multiple different mitochondrial genomes within an individual parasite (heteroplasmy) that were, however, not sufficiently divergent to represent a major source of typing error. We strongly recommend this combined nuclear and mitochondrial genotyping methodology to reveal cryptic diversity and genetic exchange in T. cruzi. The level of resolution that this mtMLST provides should greatly assist attempts to elucidate the complex interactions between parasite genotype, clinical outcome and disease distribution.

Trypanosoma cruzi, displays remarkable genetic diversity and is currently recognized as a complex of six lineages or discrete typing units (DTUs), each broadly associated with disparate ecologies and currently recognized as a complex of six lineages or discrete typing units (DTUs) [6,7]. However, the number of ancestral nuclear clades (two or three) remains enigmatic. DTU nomenclature has recently been revised by international consensus to reflect the current understanding of T. cruzi genetic diversity [8]. Several evolutionary scenarios have been proposed to account for the emergence of two hybrid lineages (TcV and TcVI) and their parental progenitors (TcII and TcIII). However, the number of ancestral nuclear clades (two or three) remains controversial [9,10].

TcI is the most abundant and widely dispersed of all T. cruzi lineages, with an ancient parental origin estimated at ~0.5–0.9 MYA [11]. The distribution of domestic TcI, propagated by domiciliated triatomine vector species, principally extends from the Amazon Basin northwards, where it is implicated as the main cause of Chagas disease in endemic areas such as Venezuela and Colombia [12,13]. Tcl is also ubiquitous in sylvatic transmission cycles throughout South America and extends into North and Central America [14,15]. Recent advances in new high resolution genotyping techniques have seen a resurgence of interest in unravelling Tcl intra-lineage diversity. In Colombia, sequencing of the mini-exon spliced leader intergenic region (SL-IR) has subdivided Tcl isolates from domestic and sylvatic transmission cycles, irrespective of geographical origin [16–18]. Other studies have demonstrated geographical clustering of Tcl strains and an ecological association between specific genotypes and Didelphis hosts [19]. Higher resolution studies exploiting multiple microsatellite markers (MLMT) also report limited gene flow between sylvatic and domestic transmission cycles manifesting as genetic diversity between Tcl isolates from sympatric sites [20,21]. In addition, unexpectedly high levels of homozygosity in multiple clones from single hosts may be indicative of recombination between similar genotypes (inbreeding) or recurrent, genome-wide, and dispersed gene conversion [20,22]. The frequency and mechanism of natural intra-Tcl genetic exchange are thus unknown, largely due to inappropriate or inadequate sampling. Evidence for such recombination is increasing and has already been documented among strains isolated from sylvatic Didelphis and Rhodnius in the Amazon Basin [23] and within a domestic/peridomestic Tcl population in Ecuador [21]. Furthermore, the generation of intra-lineage Tcl hybrids in vitro indicates that this ancestral lineage has an extant capacity for genetic exchange [24].

In kinetoplastids, the mitochondrial genome is represented by 20–50 maxicircles (20–40 kb) which, together with thousands of minicircles (0.5–10 kb), form a catenated network or kinetoplast (kDNA), comprising 20–25% of total cellular DNA [25]. Maxicircles are the functional equivalent of eukaryotic mitochondrial DNA, encoding genes for mitochondrial rRNAs and hydrophobic proteins involved in energy transduction by oxidative phosphorylation [26]. Previously, phylogenetic analyses of T. cruzi maxicircle fragments classified isolates into three mitochondrial clades A (TcI), B (TcIII, TcIV, TcV and TcVI) and C (TcII) [10,27]. To date, maxicircle typing has been principally used to examine T. cruzi inter-lineage diversity, with sequencing efforts reliant on a limited number of genes [28] and often in the absence of any comparative nuclear targets [29,30]. However, the inherent features of mitochondrial markers argue for their inclusion as principal but not solitary components of phylogenetic studies. Indeed, the caveats highlighted by other eukaryotes are especially pertinent with respect to T. cruzi. Mitochondrial introgression has been reported in North America where identical maxicircles circulate in sympatric Tcl and TcIV from sylvatic reservoirs [27] and in South America where maxicircle haplotypes are shared between TcIII and TcIV strains with highly divergent nuclear genomes [11]. However, this phenomenon has not been described among South American Tcl isolates. In addition, mitochondrial heteroplasmy, a possible confounder of phylogenetic studies, has not been examined in the coding region of the T. cruzi maxicircle but is not unexpected considering the presence of up to fifty maxicircle copies within an individual parasite.

The potential for mitochondrial DNA to reveal diversity hidden at the sub-DTU level in T. cruzi has been largely overlooked. To address this deficit, we first employed a whole genome approach to investigate the existence of maxicircle heteroplasmy and to resolve its role as a source of genotyping error. Secondly, we exploited the online availability of three complete T. cruzi maxicircle genomes [31,32] to develop a high resolution mitochondrial multilocus typing scheme (mtMLST) in order to describe Tcl intra-lineage diversity. Lastly, we investigated the extent of incongruence between mitochondrial and nuclear loci (SL-IR, GPI and 25 short tandem repeat (STR) loci) to detect incidences of genetic exchange.

Materials and Methods

Illumina Sequencing of the Sylvio X10/1 Maxicircle Genome

The maxicircle genome from Sylvio X10/1 (TcI) was sequenced at 183X coverage using Illumina HiSeq 2000 technology as part of the Sylvio X10/1 Whole Genome Shotgun project [33]. A total of 66,682 reads were generated which covered the maxicircle coding region (15,185 bp). The consensus maxicircle genome sequence
was derived from the predominant nucleotide present across multiple read alignments at each position. However, this criterion masks minor maxicircle haplotypes (evidence of heteroplasmy) by disregarding low abundance single nucleotide polymorphisms (SNPs). To assess the presence/absence of true minor SNPs, all 66,002 reads were re-aligned to the Sylvio X10/1 maxicircle genome using the alignment software SAMtools [34] and SNPs were called using the SAMtools mpileup commands. A SNP was defined as a nucleotide variant present in at least 5 independent reads (with parameters: 20X coverage; and mapping quality, 30). The final alignment was manually inspected using Tablet [35]. In parallel, ten maxicircle gene fragments, described below, were amplified by PCR and Sanger sequenced from Sylvio X10/1.

**Strains**

A panel of 32 TcI isolates was assembled for analysis (Table 1). Parasites (epimastigotes) were cultured at 28°C in RPMI-1640 liquid medium supplemented with 0.5% (w/v) tryptone, 20 mM HEPES buffer pH 7.2, 30 mM haemin, 10% (v/v) heat-inactivated fetal cell serum, 2 mM sodium glutamate, 2 mM sodium pyruvate and 25 µg/ml gentamycin (Sigma, UK) [23]. Genomic DNA was extracted using the Gentra PureGene Tissue Kit (Qiagen, UK), according to the manufacturer’s protocol. Isolates were previously characterized to DTU level using a triple-marker assay [36] and classified into seven genetic populations by microsatellite profiling [20]: North and Central America (AMNorth/Central), Venezuelan sylvatic (VENsyl), North-Eastern Brazil (BRAZNorth-East), Northern Bolivia (BOINorth), Northern Argentina (ARGNorth), Bolivian and Chilean Andes (ANDES/Bol/Chile) and Venezuelan domestic (VENdom). Genotypes for additional TcI–TcVI strains were included for comparison in selected analyses as indicated (Tables S1 and S2).

**Maxicircle Genes (mtMLST)**

Ten maxicircle gene fragments were amplified: ND4 (NADH dehydrogenase subunit 4), ND1 (NADH dehydrogenase subunit 1), ND2, ND3, ND5, ND6, CO1 (cytochrome c oxidase subunit I), CO2 (cytochrome c oxidase subunit II), CO3 (cytochrome c oxidase subunit III) and CO4 (cytochrome c oxidase subunit IV).

<table>
<thead>
<tr>
<th>Strain Code</th>
<th>Strain</th>
<th>Date of isolation</th>
<th>Host/Vector</th>
<th>Locality</th>
<th>Latitude</th>
<th>Longitude</th>
<th>Population</th>
</tr>
</thead>
<tbody>
<tr>
<td>9209</td>
<td>92090802P cl1</td>
<td>1992</td>
<td>Didelphis marsupialis</td>
<td>Georgia, USA</td>
<td>32.43</td>
<td>-83.31</td>
<td>AMNorth/Central</td>
</tr>
<tr>
<td>9307</td>
<td>93070103P cl1</td>
<td>1993</td>
<td>Didelphis marsupialis</td>
<td>Georgia, USA</td>
<td>32.43</td>
<td>-83.31</td>
<td>AMNorth/Central</td>
</tr>
<tr>
<td>DAVIS</td>
<td>DAVIS 9.90 cl1</td>
<td>1983</td>
<td>Triatoma dimidiata</td>
<td>Tegucigalpa, Honduras</td>
<td>14.08</td>
<td>-87.2</td>
<td>AMNorth/Central</td>
</tr>
<tr>
<td>OPO5</td>
<td>USAOPPOSUM cl2</td>
<td>Unknown</td>
<td>Didelphis marsupialis</td>
<td>Louisiana, USA</td>
<td>30.5</td>
<td>-91</td>
<td>AMNorth/Central</td>
</tr>
<tr>
<td>ARMA</td>
<td>USAARMA cl3</td>
<td>Unknown</td>
<td>Dasypus novemcinctus</td>
<td>Louisiana, USA</td>
<td>30.5</td>
<td>-91</td>
<td>AMNorth/Central</td>
</tr>
<tr>
<td>COT38</td>
<td>COTMA38</td>
<td>13.10.04</td>
<td>Akodon boliviensis</td>
<td>Cotopachi, Bolivia</td>
<td>-17.43</td>
<td>-66.27</td>
<td>ANDES/Bol/Chile</td>
</tr>
<tr>
<td>P234</td>
<td>P234</td>
<td>1985</td>
<td>Homo sapiens</td>
<td>Cochabamba, Bolivia</td>
<td>-17.38</td>
<td>-66.16</td>
<td>ANDES/Bol/Chile</td>
</tr>
<tr>
<td>P238</td>
<td>P238</td>
<td>1985</td>
<td>Homo sapiens</td>
<td>Cochabamba, Bolivia</td>
<td>-17.38</td>
<td>-66.16</td>
<td>ANDES/Bol/Chile</td>
</tr>
<tr>
<td>P268</td>
<td>P268</td>
<td>1987</td>
<td>Homo sapiens</td>
<td>Cochabamba, Bolivia</td>
<td>-17.38</td>
<td>-66.16</td>
<td>ANDES/Bol/Chile</td>
</tr>
<tr>
<td>PAL23</td>
<td>PALDAV2-3</td>
<td>23.03.01</td>
<td>Triatoma infestans</td>
<td>Chaco, Argentina</td>
<td>-27.133</td>
<td>-61.46</td>
<td>ARGNorth</td>
</tr>
<tr>
<td>PAL4</td>
<td>PALDA4</td>
<td>23.03.01</td>
<td>Didelphis albiventris</td>
<td>Chaco, Argentina</td>
<td>-27.133</td>
<td>-61.46</td>
<td>ARGNorth</td>
</tr>
<tr>
<td>PAL5</td>
<td>PALDA5</td>
<td>23.03.01</td>
<td>Didelphis albiventris</td>
<td>Chaco, Argentina</td>
<td>-27.133</td>
<td>-61.46</td>
<td>ARGNorth</td>
</tr>
<tr>
<td>PAL20</td>
<td>PALDA20</td>
<td>23.03.01</td>
<td>Didelphis albiventris</td>
<td>Chaco, Argentina</td>
<td>-27.133</td>
<td>-61.46</td>
<td>ARGNorth</td>
</tr>
<tr>
<td>PAL21</td>
<td>PALDA21</td>
<td>23.03.01</td>
<td>Didelphis albiventris</td>
<td>Chaco, Argentina</td>
<td>-27.133</td>
<td>-61.46</td>
<td>ARGNorth</td>
</tr>
<tr>
<td>SJ34</td>
<td>SJM34</td>
<td>07.09.04</td>
<td>Didelphis marsupialis</td>
<td>Beni, Bolivia</td>
<td>-14.81</td>
<td>-64.6</td>
<td>BOLNorth</td>
</tr>
<tr>
<td>SJ41</td>
<td>SJM41</td>
<td>09.09.04</td>
<td>Philander opossum</td>
<td>Beni, Bolivia</td>
<td>-14.81</td>
<td>-64.6</td>
<td>BOLNorth</td>
</tr>
<tr>
<td>SJ37</td>
<td>SJM37</td>
<td>09.09.04</td>
<td>Didelphis marsupialis</td>
<td>Beni, Bolivia</td>
<td>-14.81</td>
<td>-64.6</td>
<td>BOLNorth</td>
</tr>
<tr>
<td>SJ12</td>
<td>SJMC12</td>
<td>13.09.04</td>
<td>Philander opossum</td>
<td>Beni, Bolivia</td>
<td>-14.81</td>
<td>-64.6</td>
<td>BOLNorth</td>
</tr>
<tr>
<td>SJ22</td>
<td>SJM22 c1</td>
<td>06.09.04</td>
<td>Didelphis marsupialis</td>
<td>Beni, Bolivia</td>
<td>-14.81</td>
<td>-64.6</td>
<td>BOLNorth</td>
</tr>
<tr>
<td>SJ39</td>
<td>SJM39 c3</td>
<td>09.09.04</td>
<td>Didelphis marsupialis</td>
<td>Beni, Bolivia</td>
<td>-14.81</td>
<td>-64.6</td>
<td>BOLNorth</td>
</tr>
<tr>
<td>M13</td>
<td>M13</td>
<td>12.06.04</td>
<td>Didelphis marsupialis</td>
<td>Barinas, Venezuela</td>
<td>7.5</td>
<td>-71.23</td>
<td>VENsyl</td>
</tr>
<tr>
<td>M16</td>
<td>M16 cl4</td>
<td>13.06.04</td>
<td>Didelphis marsupialis</td>
<td>Barinas, Venezuela</td>
<td>7.5</td>
<td>-71.23</td>
<td>VENsyl</td>
</tr>
<tr>
<td>M18</td>
<td>M18</td>
<td>13.06.04</td>
<td>Didelphis marsupialis</td>
<td>Barinas, Venezuela</td>
<td>7.5</td>
<td>-71.23</td>
<td>VENsyl</td>
</tr>
<tr>
<td>M7</td>
<td>M7</td>
<td>14.05.04</td>
<td>Didelphis marsupialis</td>
<td>Barinas, Venezuela</td>
<td>7.5</td>
<td>-71.23</td>
<td>VENsyl</td>
</tr>
<tr>
<td>XE51</td>
<td>XE5167 cl1</td>
<td>14.09.99</td>
<td>Didelphis marsupialis</td>
<td>Para, Brazil</td>
<td>-1.71</td>
<td>-48.88</td>
<td>BRAZNorth-East</td>
</tr>
<tr>
<td>IM48</td>
<td>IM4810</td>
<td>23.04.02</td>
<td>Didelphis marsupialis</td>
<td>Manaus, Brazil</td>
<td>-3.07</td>
<td>-60.16</td>
<td>BRAZNorth-East</td>
</tr>
<tr>
<td>XE29</td>
<td>XE2929</td>
<td>10.08.88</td>
<td>Didelphis marsupialis</td>
<td>Pará, Brazil</td>
<td>-5.83</td>
<td>-48.03</td>
<td>BRAZNorth-East</td>
</tr>
<tr>
<td>B2085</td>
<td>B2085</td>
<td>03.01.91</td>
<td>Didelphis marsupialis</td>
<td>Belem, Brazil</td>
<td>-1.36</td>
<td>-48.36</td>
<td>BRAZNorth-East</td>
</tr>
<tr>
<td>1180</td>
<td>11804</td>
<td>2003</td>
<td>Homo sapiens</td>
<td>Portuguesa, Venezuela</td>
<td>9.01</td>
<td>-69.29</td>
<td>VENDom</td>
</tr>
<tr>
<td>1171</td>
<td>11713</td>
<td>2003</td>
<td>Homo sapiens</td>
<td>Lara, Venezuela</td>
<td>10.04</td>
<td>-69.32</td>
<td>VENDom</td>
</tr>
<tr>
<td>9354</td>
<td>9354</td>
<td>1999</td>
<td>Homo sapiens</td>
<td>Sucre, Venezuela</td>
<td>10.46</td>
<td>-63.61</td>
<td>VENDom</td>
</tr>
<tr>
<td>1154</td>
<td>11541</td>
<td>2003</td>
<td>Homo sapiens</td>
<td>Merida, Venezuela</td>
<td>8.59</td>
<td>-71.23</td>
<td>VENDom</td>
</tr>
</tbody>
</table>

*Strain code corresponds to labels on Figure 3 and descriptions in text. doi:10.1371/journal.pntd.0001584.s001*
and NED, PET and VIC (Applied Biosystems, UK). Allele sizes
label the forward primers: 6-FAM and TET (Proligo, Germany)
of each primer, 1 U Taq polymerase (Bioline, UK), 0.2 mM dNTPs
(New England Biolabs, UK), 10 pmol of each primer, 1 U Taq polymerase (Bioline, UK) and
10–100 ng of genomic DNA. PCR reactions were performed with
an initial denaturation step of 3 minutes at 94°C, followed by 30
amplification cycles (94°C for 30 seconds, 50°C for 30 seconds, 72°C for 30 seconds) and a final elongation step at 72°C for ten
minutes. PCR products were purified using QIAquick PCR extraction kits (Qiagen, UK) according to the manufacturer’s protocol.

Nuclear Genes
The mini-exon spliced leader intergenic region (SL-IR) and glucose-6-phosphate isomerase (GPI) were amplified as previously described by Souto et al. (1996) [39] and Lewis et al. (2009) [36], respectively. PCR products were visualized in 1.5% agarose gels and if necessary purified using QIAquick PCR and gel extraction kits (Qiagen, UK) to remove non-specific products. Bi-directional sequencing was performed for both nuclear and maxicircle targets using the BigDye® Terminator v3.1 Cycle Sequencing Kit (Applied Biosystems, UK) according to the manufacturer’s protocol. Maxicircle PCR products were sequenced using the relevant PCR primers described in Table 2. Nuclear amplicons were sequenced using their respective PCR primers. When ambiguous sequences were obtained, PCR products were cloned into the pGEM® - T Easy Vector System I (Promega, USA) according to the manufacturer’s instructions, and transformed into XL1-Blue E. coli (Agilent Technologies, UK), prior to colony PCR and re-sequencing. For strains that produced incongruent nuclear and maxicircle phylogenetic signals, PCR and sequencing reactions were replicated twice using DNA derived from two independent genomic DNA extractions.

Microsatellite Loci
Data from 25 previously described microsatellite loci [20], distributed among ten chromosomes [40], were included for analysis. Loci were selected from a wider panel of 48 microsatellite loci based on their level of Tcl intra-lineage resolution. In addition, these 25 microsatellite loci were amplified across eight new unpublished biological clones (M16 c14, SJM22 c11, SJM39 c13, USAARMA c13, USAOPOSSUM c12, 92000802P c11, 93070103P c11 and DAVIS 9.90 c11). Primers and binding sites are listed in Table S3. The following reaction conditions were implemented across all loci: a denaturation step of 4 minutes at 95°C, then 30 amplification cycles (95°C for 20 seconds, 57°C for 20 seconds, 72°C for 20 seconds) and a final elongation step at 72°C for 20 minutes. Amplifications were achieved in a final volume of 10 µl containing: 1 x ThermoPol Reaction Buffer (New England Biolabs, UK), 4 mM MgCl₂, 34 µM dNTPs, 0.75 pmol of each primer, 1 U Taq polymerase (New England Biolabs, UK) and 1 ng of genomic DNA. Five fluorescent dyes were used to label the forward primers: 6-FAM and TET (Proligo, Germany) and NED, PET and VIC (Applied Biosystems, UK). Allele sizes were determined using an automated capillary sequencer (AB3730, Applied Biosystems, UK), in conjunction with a fluorescently tagged size standard, and were manually checked for errors. All isolates were typed “blind” to control for user bias.

Phylogenetic Analysis of Nuclear Loci
Pair-wise distances (DAS) between microsatellite genotypes for individual samples were calculated in MICROSAT v1.5d [41] under the infinite-alleles model (IAM). To accommodate multi-allelic genotypes (≥3 alleles per locus), a script was written in Microsoft Visual Basic to generate random multiple diploid re-samplings of each multilocus profile (software available on request). A final pair-wise distance matrix was derived from the mean of each re-sampled dataset and used to construct a Neighbour-Joining phylogenetic tree in PHYLIPT v3.67 [42]. Majority rule consensus analysis of 10,000 bootstrap trees was performed in PHYLIPT v3.67 by combining 100 bootstraps created in MICROSAT v1.5d, each drawn from 100 respective randomly re-sampled datasets.

Nucleotide sequences were assembled manually in BioEdit v7.0.9.0 sequence alignment editor software (Ibis Biosciences, USA) [43] and unambiguous consensus sequences were produced for each isolate. Heterozygous SNPs were identified by the presence of two coincident peaks at the same locus (‘split peaks’), verified in forward and reverse sequences and scored according to the one-letter nomenclature for nucleotides from the International Union of Pure and Applied Chemistry (IUPAC). For both nuclear genotypes (SL-IR and GPI), edited sequences were used to generate Neighbour-Joining trees based on the Kimura-2 parameter model in MEGA v5 [44]. Bootstrap support for clade topologies was estimated following the generation of 1000 pseudo-replicate datasets. Once both trees were visualized independently to confirm congruent topologies, nucleotide SNPs were re-coded numerically and concatenated with microsatellite data (see Dataset S1). DAS values were calculated for the concatenated dataset as described above and used to generate a single Neighbour-Joining phylogenetic tree encompassing all nuclear genetic diversity. Nucleotide sequences for GPI and the SL-IR are available from GenBank under the accession numbers JQ581371–JQ581402 and JQ581401–JQ581512, respectively.

Phylogenetic Analysis of Maxicircle Genes
Nucleotide data were assembled manually as described for nuclear loci. For each isolate, maxicircle sequences were concatenated according to their structural arrangement (12S rRNA, 9S rRNA, CYT b, MURF1, ND1, COII, ND4 and ND5) and in the correct coding direction (alignment available on request). Nucleotide sequences for all ten gene fragments are available from GenBank under the accession numbers listed in Table 2. Phylogenies were inferred using Maximum-Likelihood (ML) implemented in PhyML (4 substitution rate categories) [45]. The best-fit model of nucleotide substitution was selected from 50 models and its significance evaluated according to the Akaike Information Criterion (AIC) in jMODELTEST 1.0. [46]. The best model selected for this dataset was GTR+I+G. Bootstrap support for clade topologies was estimated following the generation of 1000 pseudo-replicate datasets. Bayesian phylogenetic analysis was performed using MrBAYES v3.1 [47] (settings according to jMODELTEST 1.0). Five independent analyses were run using a random starting tree with three heated chains and one cold chain over 10 million generations with sampling every 10 simulations (25% burn-in). Shimodaira-Hasegawa likelihood tests (SH tests) [48] were implemented in PAML v4.4 [49] to statistically
<table>
<thead>
<tr>
<th>Gene Fragment</th>
<th>Genome Position</th>
<th>Primer Name</th>
<th>Primer Sequence (5’→3’)</th>
<th>Annealing Temp. (°C)</th>
<th>Amplicon Size (bp)</th>
<th>Sequence Start 5’</th>
<th>Sequence Start 3’</th>
<th>Sequenced Fragment (bp)</th>
<th>Accession Numbers</th>
</tr>
</thead>
<tbody>
<tr>
<td>12S rRNA</td>
<td>639–901</td>
<td>12S Fwd</td>
<td>GTTATTAATGGCGTTGCTAAGAA</td>
<td>50</td>
<td>299</td>
<td>GTCTAAGA</td>
<td>TACGTATT</td>
<td>263</td>
<td>JQ581254–JQ581292</td>
</tr>
<tr>
<td></td>
<td></td>
<td>12S Rvs</td>
<td>GCCCAATCAACACATCAACAA</td>
<td>(19)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>9S rRNA</td>
<td>1077–1309</td>
<td>9S Fwd</td>
<td>TGCAATTCTGTAGTTGGGTTGA</td>
<td>50</td>
<td>302</td>
<td>TAAAAATCG</td>
<td>TATTATTA</td>
<td>233</td>
<td>JQ581215–JQ581253</td>
</tr>
<tr>
<td></td>
<td></td>
<td>9S Rvs</td>
<td>TCCACACACATATATACACT</td>
<td>(22)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CYT b</td>
<td>4126–4733</td>
<td>Sp18 Fwd</td>
<td>GACAGATGAGAGAGGGAGAGAG</td>
<td>50</td>
<td>717</td>
<td>TTGTCTT</td>
<td>TAATAYCA</td>
<td>608</td>
<td>JQ581332–JQ581370</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Sp18 Rvs</td>
<td>CAAAACCTACACACAAAAGCTTCTG</td>
<td>(24)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Murf1a</td>
<td>6011–6393</td>
<td>Murf1a Fwd</td>
<td>AAGGCRATGGGRATAGWRCCTATAC</td>
<td>50</td>
<td>482</td>
<td>ACTAAGYA</td>
<td>ACTTTYTA</td>
<td>383</td>
<td>JQ581403–JQ581441</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Murf1a Rvs</td>
<td>TGGAAATCTTATATACAGATTG</td>
<td>(26)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Murf1b</td>
<td>6528–6900</td>
<td>Murf1b Fwd</td>
<td>ACMCCATCCATCTCTCR</td>
<td>50</td>
<td>423</td>
<td>CAAAATTT</td>
<td>GGATTTAT</td>
<td>373</td>
<td>JQ581442–JQ581480</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Murf1b Rvs</td>
<td>CTTTGGATTTTGTGTGGATTTTGTG</td>
<td>(25)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ND1</td>
<td>7643–8011</td>
<td>ND1 Fwd</td>
<td>GCACCTTCTGAATATAGAAGA</td>
<td>50</td>
<td>400</td>
<td>TCGAAAAA</td>
<td>TGTTAGC</td>
<td>369</td>
<td>JQ581059–JQ581097</td>
</tr>
<tr>
<td></td>
<td></td>
<td>ND1 Rvs</td>
<td>TTAACCTTATAGGAGTTGAGCC</td>
<td>(25)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>COII</td>
<td>8194–8610</td>
<td>COII Fwd</td>
<td>GTTATACATTCTTGTGGTGGTG</td>
<td>50</td>
<td>560</td>
<td>CTTCTAC</td>
<td>ACCTRCCY</td>
<td>417</td>
<td>JQ581293–JQ581331</td>
</tr>
<tr>
<td></td>
<td></td>
<td>COII Rvs</td>
<td>CAAAATGAGCATACATACATG</td>
<td>(22)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ND4</td>
<td>12153–12392</td>
<td>ND4 Fwd</td>
<td>TTTTGGAGCTCTATTATACCCCA</td>
<td>50</td>
<td>302</td>
<td>AATATTA</td>
<td>CCGTGYRC</td>
<td>240</td>
<td>JQ581098–JQ581136</td>
</tr>
<tr>
<td></td>
<td></td>
<td>ND4 Rvs</td>
<td>CTCAACACATCCTTCCGTTT</td>
<td>(21)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ND5a</td>
<td>13829–14250</td>
<td>ND5a Fwd</td>
<td>TATGATCAACCTTCTTTCGTC</td>
<td>50</td>
<td>503</td>
<td>GTACATAY</td>
<td>TTYTYGTA</td>
<td>422</td>
<td>JQ581137–JQ581175</td>
</tr>
<tr>
<td></td>
<td></td>
<td>ND5a Rvs</td>
<td>GTCTCCATGTGCTACATGCGG</td>
<td>(19)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ND5b</td>
<td>14274–14640</td>
<td>ND5b Fwd</td>
<td>ARAGTACACACCGGAGTCAYCA</td>
<td>50</td>
<td>444</td>
<td>TGATTRCC</td>
<td>GYARACCA</td>
<td>367</td>
<td>JQ581176–JQ581214</td>
</tr>
<tr>
<td></td>
<td></td>
<td>ND5b Rvs</td>
<td>CTGTGAYACATACACCAACAA</td>
<td>(21)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

aGenome position according to the TcI Sylvio X10/1 reference maxicircle genome [33].
bAmplicon size according to TcI Sylvio X10/1. Indels in other strains may cause size variation.
cSequence length according to TcI Sylvio X10/1. Indels in other strains may cause length variation.
doi:10.1371/journal.pntd.0001584.t002
evaluate incongruencies between alternative tree topologies derived from the mitochondrial and nuclear data.

Results

Maxicircle Heteroplasmy

Across the 15,185 bp of the Sylvio X10/1 maxicircle coding region a total of 74 SNPs were identified among eight genes (12S rRNA, 9S rRNA, MURF5, CYT b, MURF1, MURF2, CR4 and ND4) and three intergenic regions (between 12S rRNA and 9S rRNA, between 9S rRNA and ND4 and between CR4 and ND4, respectively) (Figure 2 and Table S4). Average read depth for each SNP site was 163. At heterozygous sites, the minor nucleotide was present among an average of 12.2% (±9.1%) of sequence reads. In each gene, SNPs were clustered often <5 bp apart in pairs and triplets. The most common mutations were transversions from A→T (14/74), T→A (10/74), T→G (7/74) and G→T (6/74) and transitions from A→G (13/74). SNPs were bi-variable at all sites. The presence of different contiguous SNPs distributed across separate sequencing reads at overlapping positions suggests the occurrence of at least two minor maxicircle templates within the same sample. However, the short average length of Illumina reads (~100 bp) prohibits the full reconstruction of minor maxicircle sequence types. No evidence of heterozygosity was observed in any of the ten maxicircle Sanger sequences (from the mtMLST scheme) that covered the corresponding areas of heteroplasmy identified in Sylvio X10/1, which is consistent with the low sensitivity of this method.
Maxicircle Genes (mtMLST)
Degenerate primers were designed by reference to complete TcI, TcII and TcVI maxicircle genomes. Ten gene fragments from eight maxicircle coding regions were selected in order to sample genetic diversity present across the whole T. cruzi maxicircle. For two genes (MURF1 and ND5) two fragments were selected from each coding region to examine intra-gene variation. Reliable PCR amplification of all ten maxicircle fragments was first confirmed using a panel of T. cruzi reference strains from each DTU (see Figure 1).

The maxicircle gene targets were then sequenced across the TcI panel (Table 1) and seven additional TcIII/TcIV strains (Table S2). Relative uniform substitution rates were observed among all genes (gamma shape parameter $\alpha = 0.8121$, based on the GTR+$\Gamma$+G model). For each TcI isolate, gene fragments were concatenated according to their structural position and assembled into a 3686 bp alignment. Twenty-two unique haplotypes were identified from a total of 355 variable sites (~9.6% sequence diversity). No evidence of heterozygosity ('split peaks') was observed.

Maximum-Likelihood (Figure 3, right) and Bayesian phylogenies were both constructed from the concatenated maxicircle data. No statistically-supported incongruence was observed between the two topologies (Bayesian tree $L = -6770.21$, ML tree $L = -6768.85$, $P = 0.428$). The presence of at least three incongruent haplotypes (see below) precludes the accurate clustering of their respective populations (AM$_{North/Cen}$, VEN$_{dom}$ and BRAZ$_{North-East}$). However, phylogenetic analysis does resolve two well-supported clades corresponding to VEN$_{silv}$ and ARG$_{North}$, and ANDES$_{Bol/Chile}$ (90.8%/1.0 and 100%/1.0, respectively). Once the two TcIV-type maxicircles were excluded from analysis, the mtMLST was re-evaluated with respect to intra-TcI discriminatory power. One hundred SNPs were identified among 3681 bp (~2.7% sequence diversity), corresponding to twenty maxicircle haplotypes. Both Bayesian and Maximum-Likelihood topologies were congruent with those constructed previously for the entire TcI isolate panel.

Nuclear Loci
The resolutive power of the mtMLST scheme was evaluated by comparison to current markers used to investigate TcI intra-DTU nuclear diversity, specifically, a housekeeping gene ($GPI$), a non-coding multi-copy intergenic region (SL-IR) and a MLMT panel of 25 loci. Sequences for $GPI$ were obtained for 32 T. cruzi isolates (Table 1) and assembled into a gap-free alignment of 921 nucleotides. Of the 921 bp, a total of 911 invariable sites and 10 polymorphic sites were identified (~1.1% sequence diversity). A 350 bp alignment corresponding to the SL-IR was generated for the same panel of samples. Strains from two populations (5/6 BOL$_{North}$ and 4/4 ANDES$_{Bol/Chile}$) presented sequences with multiple ambiguous base calls due to the presence of a GT$_{A}$ microsatellite at positions 14–24. For these nine isolates, haplotypes were determined by sequencing four cloned PCR products to derive a consensus sequence. In the 350 bp alignment, 323 conserved sites and 36 polymorphic sites were observed (~10.3% sequence diversity). All samples were also typed at 25 polymorphic microsatellite loci yielding a total of 1612 alleles. The majority of strains presented one or two alleles at each locus. Multiple alleles (~2) were observed at a small proportion of loci (1.5%).

Individual Neighbour-Joining trees were re-constructed for $GPI$, SL-IR and the MLMT data. No well-supported sub-DTU level clades were recovered using $GPI$ sequences. The SL-IR phylogeny resolved two populations (VEN$_{silv}$, and ARG$_{North}$) with strong statistical support (85% and 99%, respectively; data not shown). Three major clades were identified by MLMT (VEN$_{dom}$, ARG$_{North}$ and ANDES$_{Bol/Chile}$) with good bootstrap support (72.6%, 90.3% and 98.4%, respectively; data not shown). There was no bootstrap-supported incongruence between the three nuclear tree topologies. This justified their concatenation and these data were re-coded and analyzed in a single distance-based phylogeny (independent of mutation rate heterogeneity) (Figure 3, left and Dataset S1). The concatenated nuclear tree recovered three well supported clades corresponding to TcI populations (VEN$_{silv}$, ARG$_{North}$ and ANDES$_{Bol/Chile}$) with strong bootstrap support (96%, 100% and 77.9%, respectively, Figure 3). Isolates belonging to the VEN$_{dom}$ population remained grouped together but with a minor reduction in bootstrap values (64.8%), compared to the MLMT tree. In addition, the concatenated tree also subdivided BOL$_{North}$ into two well defined sympatric clades each containing three isolates (99.8% and 82.2%). No nuclear targets (either individually or concatenated) were able to reliably identify AM$_{North/Cen}$ or BRAZ$_{North-East}$ as discrete clusters. However, AM$_{North/Cen}$ was more closely related to VEN$_{dom}$ than any other population by MLMT (90.2%), the SL-IR (99%) and the concatenated nuclear tree (100%).

Figure 2. Distribution of seventy-four heteroplasmic sites across the 15,185 bp Sylvio X10/1 maxicircle genome (schematic shows linearized maxicircle).

66,882 sequencing reads covering the Sylvio X10/1 maxicircle were generated using Illumina HiSeq 2000 technology as part of the Sylvio X10/1 Whole Genome Shotgun project. Multiple reads were re-aligned to the maxicircle genome and SNPs were identified if a nucleotide variant was present in at least five independent reads. Bars represent the abundance of major (reference nucleotide) and minor bases among multiple reads at each position. All SNPs are bi-variable. At some overlapping positions, different contiguous SNPs are distributed among separate sequencing reads. These observations suggest the occurrence of at least two additional maxicircle genomes at a ~10-fold lower abundance compared to the consensus genome. Red stars denote gene fragments used in the mtMLST scheme.
Nuclear-Mitochondrial Incongruence

Comparison of the mitochondrial and nuclear phylogenies revealed clear incongruence at multiple scales. The nuclear topology was a significantly worse model to fit the maxicircle data ($L = 27008.72$, mtMLST ML tree $L = 26554.50$, $P = < 0.001$). Three individual isolates had unambiguously different phylogenetic positions between the nuclear and mitochondrial datasets: 9307, 9354 and IM48 (Figure 3). The maxicircle sequences from 9307, a sylvatic TcI AM_North/Cen strain, and 9354, a human TcI strain from VEN_dom, were divergent from all other TcI strains. Comparison with sequences from other DTUs indicates that the maxicircle from 9307 was most closely related to those found in TcIV samples from North America (92122) (100%) while 9354 shared its mitochondrial haplotype with TcIV and TcIII strains from neighbouring areas of Venezuela, Bolivia and Colombia. Divergent maxicircle haplotypes at the intra-DTU level are also observed in BRAZ_North-East (IM48) and AM_North/Cen (ARMA and OPOS). Another incidence of nuclear-mitochondrial incongruence is demonstrated by the paraphyletic grouping of ARG_North among a subset of BOl_North isolates in the maxicircle tree, compared to its monophyletic placement in the nuclear phylogeny.

doi:10.1371/journal.pntd.0001584.g003

Figure 3. Unrooted Neighbour-Joining tree based on $D_{AS}$ values from nuclear loci (left) and Maximum-Likelihood tree from concatenated maxicircle sequences (right) showing TcI population structure across the Americas. A panel of 32 TcI isolates from seven nuclear populations was assembled for analysis. Origin of individual strains is shown on the map by small red circles. Large red circles correspond to multiple samples, isolated from the same geographical area. Branch colours indicate strain population. The nuclear tree was constructed from concatenated polymorphisms present within the SL-IR, GPI and 25 microsatellite loci. $D_{AS}$ values were calculated as the mean across 1000 random diploid re-samplings of the dataset and those greater than 70% are shown on major clades. A Maximum-Likelihood topology was assembled from concatenated maxicircle sequences. Branches show equivalent bootstraps and posterior probabilities from consensus Maximum-Likelihood (1000 replicates) and Bayesian topologies, respectively. The maxicircle topology is rooted against additional outgroup strains from TcII and TcIV. The blue and red circles on branches represent inter-lineage introgression events. The blue circle indicates that the maxicircle in a sylvatic TcI isolate from AM_North/Cen is most closely related to the maxicircle found in TcIV samples from the same area. The red circle shows that the maxicircle haplotype in a human VEN_dom strain is the same as those in TcII and TcIV isolates from neighbouring areas of Venezuela, Bolivia and Colombia. Divergent maxicircle haplotypes at the intra-DTU level are also observed in BRAZ_North-East (IM48) and AM_North/Cen (ARMA and OPOS). Another incidence of nuclear-mitochondrial incongruence is demonstrated by the paraphyletic grouping of ARG_North among a subset of BOl_North isolates in the maxicircle tree, compared to its monophyletic placement in the nuclear phylogeny.

Mitochondrial Introgression in T. cruzi

Nuclear-Mitochondrial Incongruence

Comparison of the mitochondrial and nuclear phylogenies revealed clear incongruence at multiple scales. The nuclear topology was a significantly worse model to fit the maxicircle data (nuclear tree $L = -7008.72$, mtMLST ML tree $L = -6554.50$, $P<0.001$). Three individual isolates had unambiguously different phylogenetic positions between the nuclear and mitochondrial datasets: 9307, 9354 and IM48 (Figure 3). The maxicircle sequences from 9307, a sylvatic TcI AM_North/Cen strain, and 9354, a human TcI strain from VEN_doms, were divergent from all other TcI strains. Comparison with sequences from other DTUs indicates that the maxicircle from 9307 was most closely related to those found in TcIV samples from North America (92122) (100%) while 9354 shared its mitochondrial haplotype with TcIV and TcIII strains from neighbouring areas of Venezuela, Bolivia and Colombia (ERA, 10R26, X106, Sairi3 and CM17) (97.8%/0.9). IM48 from BRAZ_North-East also had a distinct maxicircle haplotype that formed a long branch separated from the other members of this population whereas for nuclear data all BRAZ_North-East isolates, including IM48, clearly grouped together.

To test whether inclusion of these isolates could explain the overall incongruence, the SH analysis was repeated for alternative nuclear vs. mitochondrial topologies with each of these strains excluded individually and then collectively. In all cases, statistically significant incongruence persisted (no 9307 $P = 0.004$, no 9354 $P = 0.002$, no IM48 $P < 0.001$ and without all three $P = 0.008$). This indicated that mitochondrial introgression was generally pervasive in the TcI panel beyond these three isolates. For example, ARG_North samples, which formed a homogeneous monophyletic clade that was most closely related to ANDES_PanChla by nuclear data, grouped paraphyletically amongst subsets of BOl_North strains in the maxicircle tree. In addition, BRAZ_North-East is grouped with one of the BOl_North clades in the nuclear tree, but receives a basally diverging position in the maxicircle phylogeny. In agreement with
the nuclear data, AM\textsubscript{North/Cen} was most closely related to VEN\textsubscript{dom}. However, two isolates from AM\textsubscript{North/Cen} (ARMA and OPOS) displayed an unexpected level of maxicircle diversity and are grouped separately with strong bootstrap support (96.6%/1.0).

**Discussion**

Elucidating the complex epidemiology, phylogeography and taxonomy of *T. cruzi* requires a clear understanding of the parasite’s genetic diversity [4]. One objective of this study was to develop the first mitochondrial (maxicircle) multilocus sequence typing scheme (mtMLST) to investigate *T. cruzi* intra-lineage diversity and to critically assess its resolutive power compared to the current repertoire of phylogenetic markers.

The presence of intra-strain maxicircle diversity within Sylvio X10/1 is the first demonstration of heteroplasmacy in the coding region of a *T. cruzi* maxicircle genome. Seventy-four variable sites were identified by read depth analysis of Illumina sequence data but undetected by conventional Sanger sequencing. These SNPs indicate the occurrence of at least two additional maxicircle genomes, present at a ~10-fold lower abundance compared to the consensus published Sylvio X10 maxicircle genome [32]. Most heteroplasmic SNPs were linked. This may indicate an older most recent common ancestor (MRCA) between the major and minor maxicircles than that expected to have emerged in culture post-cloning. Thus these minor maxicircle classes more likely represent heteroplasmacy within a single parasite than within a subpopulation of cells. Furthermore, the presence of SNPs <3 bp apart on contiguous sequence reads may have non-synonymous coding implications, although their relative rarity, and a lack of indels suggest that minority and majority maxicircle variants would not differ phenotypically. Finally, the presence of heteroplasmacy at less than 0.5% of sites indicates it is unlikely to represent a major source of typing error when using maxicircle Sanger sequencing to characterize isolates.

Several factors are likely to contribute to mitochondrial heteroplasmacy. Mutation in length or nucleotide composition and/or bi-parental inheritance in genetic exchange events are both exacerbated by differential replication rates and inequitable cytoplasmic segregation of mitochondrial genomes during mitosis [50,51]. In kinetoplastids, maxicircle intra-clone diversity in the non-coding region was previously reported in both *T. cruzi* [31] and *Leishmania major* [52,53]. In addition, an earlier study attributed a change in *T. cruzi* maxicircle gene repertoire (elimination of one of two heteroplasmic ND7 amplicons) to sub-culture [54]. However, biologically cloned samples were not used and the possibility of a mixed infection was excluded on the basis of only four microsatellite loci. Sylvio X10/1 (a biological clone produced by micromanipulation) was first isolated from a Brazilian patient in 1979 [55] and has been in intermittent sub-culture ever since. The retention of minor maxicircle classes in Sylvio X10/1 for over thirty years suggests that a heteroplasmic state in *T. cruzi* is naturally sustained.

The observations that *T. cruzi* mitochondrial heteroplasmacy is not present at sufficient levels to adversely disrupt phylogenetic reconstructions stimulated the development of the mtMLST scheme and its assessment against traditional nuclear targets. Initially, three types of nuclear marker were evaluated, each characterized by different rates of evolution. Unsurprisingly *GPI* was highly conserved across TcI and lacked sufficient resolution to discriminate between isolates. The slow accumulation of point mutations at housekeeping loci, which are generally under purifying selection, renders these targets more appropriate to describe inter-DTU variation. Thus they are valuable candidates for inclusion in traditional nuclear MLST schemes [56]. The mini-cxom SL-IR is widely used as a TcI taxonomic marker in view of its heterogeneity and ease of amplification [57]. In this study, SL-IR variability manifested as a ten-fold increase in sequence diversity as compared to that of *GPI* and supported the robust delineation of two nuclear populations (VEN\textsubscript{dom} and ARG\textsubscript{north}). However, there were several caveats associated with the SL-IR, notably the presence of multiple tandemly-repeated copies with undefined chromosomal orthology between strains [58]. Previous attempts to estimate the level of intra-isolate SL-IR diversity have reported >96% homology between copies [19]. However, only ten clones were sequenced from each sample, representing less than 10% of the ~200 copies present per genome. Recent observations of substantial variation in gene copy number and chromosomal arrangement between *T. cruzi* strains further discourage the use of such targets for taxonomy [59]. In addition, numerous indels in the SL-IR prevent the sequencing of a suitable outgroup [39] and multiple ambiguous alignments, introduced by the microsatellite region, can disrupt phylogenetic signals [60]. Ultimately both *GPI* and the SL-IR suffer from the same fundamental criticism that single genes are inadequate to infer the overall phylogeny of an entire species [61]. Recombination, gene conversion and concerted evolution have all contributed to the genealogical history of *T. cruzi* [62] but remain undetectable using single loci.

The 25 microsatellite loci afforded the highest level of resolution from an individual set of markers, defining three statistically-supported groupings (VEN\textsubscript{dom}, ARG\textsubscript{north} and ANDES\textsubscript{Bol/Chile}). Their superior performance compared to *GPI* and the SL-IR is expected considering microsatellites are neutrally-evolving, co-dominant and hypervariable with mutation rates several orders of magnitude higher than protein-coding genes [63]. However, the use of these markers is not devoid of limitations. Most importantly, microsatellites are particularly sensitive to homoplasy, a situation where two alleles are identical in sequence but not descent, and thus fail to discriminate between closely related but evolutionarily distinct strains [64]. The three nuclear markers (*GPI*, SL-IR and microsatellites) were concatenated based on the assumption that no robust incongruence was observed between individual phylogenetic trees. However, concatenating these data did not have a significant additive effect on the level of resolution, with just three populations (VEN\textsubscript{dom}, ARG\textsubscript{north} and ANDES\textsubscript{Bol/Chile}) emerging as well-supported groups. Importantly this dataset did reveal a subdivision in the BOL\textsubscript{North} group, which went undetected by all individual nuclear markers.

Gross incongruence between the mtMLST and nuclear phylogenies revealed two incidences of inter-DTU mitochondrial introgression, indicative of multiple genetic exchange events in *T. cruzi*. Introgression was detected in North America, where identical maxicircles were observed in sylvatic TcI and TcIV isolates. A 1.25 kb fragment (*COII-ND1*) of this TcIV maxicircle haplotype has been previously described in other TcI samples from the US states of Georgia and Florida [11,27]. On the basis of the limited nuclear loci examined, and in line with previous work [27], only TcI derived nuclear genetic material appears to have been retained in these hybrids. The genetic disparity between North and South American TcIV isolates, coupled with their geographical and ecological isolation [65], implies that this event most likely occurred in North/Central America. A second, independent novel mitochondrial introgression event was identified in a Venezuelan clinical isolate. This TcI strain (9354) shares its maxicircle haplotype with a subset of human and sylvatic TcIV and TcIII isolates from Bolivia, Venezuela and Colombia, consistent with a local and possibly recent origin. Presumably
TcIV, a known secondary agent of human Chagas disease in Venezuela, is a more likely donor candidate than TcIII, which is largely absent from domestic transmission cycles [4]. Nonetheless, evidence of homogeneous maxicircle sequences in multiple, geographically dispersed isolates from different transmission cycles implies the occurrence of several genetic exchange events. It is conceivable that the TcIV/TcIII-type maxicircle sampled in this study is a relic from a TcI antecedent, supporting a common ancestry between TcI, TcIII and TcIV [9]. Alternatively, this haplotype may have originated from a TcIV or TcIII strain and its distribution reflects a recent unidirectional backcrossing event into TcI. Introgression is a more parsimonious explanation than the retention of ancestral polymorphisms through incomplete lineage sorting, particularly in areas of sympathy or parapatry among DTUs [66]. However, the historical diversification of TcI [67] and TcIII [68–70], driven by disparate ecological niches [71], and the current separation between most arboreal and terrestrial transmission cycles of TcIV and TcIII, respectively, challenge the likelihood of secondary contact between these lineages, a prerequisite of introgressive hybridization. Resolving the donor DTU of this event is complicated by the presence of indistinguishable mitochondrial sequences and paradoxically divergent nuclear genes in TcIII and TcIV isolates. It is unclear whether this results from a mechanism acting to homogenize maxicircles while allowing nuclear genes to slowly deviate [11] (unlikely), repeated and recurrent backcrossing (more likely), or merely reflects the relative paucity of available TcIV and TcIII genotypes for comparison (a certainty).

Regardless of the underlying mechanisms, it is clear that genetic exchange continues to influence the natural population structure of T. cruzi TcI. In this study, the failure to detect reciprocal transfer of nuclear DNA using an array of loci readily demonstrates the importance of adopting an integrative approach, complementing traditional nuclear markers with multiple mitochondrial targets. In the absence of comparative genomics, it is impossible to establish whether mitochondrial introgression is entirely independent of nuclear recombination.

Another advantage of the mtMLST scheme is its ability to reveal cryptic sub-DTU diversity. The significantly different evolutionary histories of the nuclear and maxicircle genes from members of BOI_{North} and ARG_{North} are consistent with intra-lineage recombination. The low levels of diversity observed within this incongruent maxicircle clade are indicative of recent and possibly multiple exchange events. In addition, two divergent maxicircles from AMN_{North/Cca} have also exposed a level of diversity that conflicts with earlier reports of reduced genetic differentiation in this group resulting from their recent biogeographical expansion [18,72]. Furthermore, the incongruent basal phylogenetic position of most of BRAZ_{North-East} in the maxicircle tree as well as the presence of another divergent maxicircle in one isolate (IM48) from this population highlights the extent to which intra-lineage diversity can be neglected by other genotyping methods. The phylogenetic placement of IM48 suggests it may be the product of an intra-TcI introgression event. However, IM48 is also a geographical outlier within the BRAZ_{North-East} population and it is difficult to determine the origin of this maxicircle haplotype in the absence of additional isolates from West-Central Amazonia.

The mechanisms governing maxicircle genetic exchange and the origins of heteroplasmy observed in Sylvio X10/1 are debatable. Currently, all reported maxicircle inheritance in natural [11] and experimental T. cruzi hybrids [24] is uniparental. However, the demonstration of heteroplasmy in this study suggests that, following genetic exchange, any minor maxicircle genotypes may be undetectable using conventional sequencing techniques. In addition, evidence of bi-parental transmission of both maxicircles [73,74] and minicircles [75] in experimentally-derived T. brucei hybrids indicates that this phenomenon can occur in kinetoplastids as a result of recombination. The mechanism of genetic exchange in T. cruzi [24] differs from meiosis, which is observed in T. brucei [73,76]. Current data suggest in vivo recombination in T. cruzi may be analogous to the parasexual cycle of Candida albicans where nuclear fusion creates a tetraploid intermediate, followed by genome erosion and reversion to aneuploidy [24,77,78]. It is not implausible to suggest that the process of cell fusion and nuclear re-arrangement may be accompanied by asymmetrical kinetoplast distribution to progeny cells. Furthermore, the sequence redundancy observed among minicircle guide RNAs has been postulated to allow biparental inheritance to occur with no detrimental consequences to mitochondrial RNA editing and hybrid viability [79].

Most importantly, the phenotypic implications of mitochondrial heteroplasmy and introgression in T. cruzi are unknown. Maxicircles play a fundamental role in parasite metabolism and development in the triatomine bug vector. Therefore the relationship between genetic recombination and phenotypic heterogeneity may have important implications for disease epidemiology. mtMLST presents a valuable new strategy to detect directional gene flow and examine the dispersal history of T. cruzi at the transmission cycle level. Furthermore, mtMLST is an excellent tool to identify genetic exchange between closely related isolates in conjunction with nuclear MLMT data. By adopting a combined nuclear and mitochondrial approach, one can simultaneously address local, epidemiologically important hypotheses as well as robustly identify parasite mating systems. Thus in combination with adequate spatio-temporal sampling, we strongly recommend this methodology as an alternative to exclusively nuclear or mitochondrial population genetic studies in future work with medically important trypanosomes. Finally, the level of resolution that the mtMLST method provides should greatly facilitate attempts to elucidate the relationship between specific parasite genotypes and phenotypic traits relating to Chagas disease pathology.

Supporting Information

Table S1 Panel of reference strains from the six T. cruzi DTUs.

Table S2 Additional T. cruzi TcIII and TcIV isolates used in selected analyses.

Table S3 Microsatellite loci and primer sequences.

Table S4 Heteroplasmic sites in the Sylvio X10/1 maxicircle genome.

Dataset S1 Concatenated nuclear dataset spreadsheet. Individual Neighbour-Joining trees were constructed for both nuclear genes (SL-1R and GP1) and the 25 microsatellite loci. Once all trees were visualized independently to confirm congruent topologies, nuclear SNPs were re-coded numerically and concatenated with microsatellite data in this spreadsheet. Delta values were calculated for this concatenated dataset and used to generate a single Neighbour-Joining tree encompassing all nuclear genetic diversity.

www.plosntds.org 10 April 2012 | Volume 6 | Issue 4 | e1584
Acknowledgments

The authors thank Maikell Segovia, Anahi Alberti and Patricio Diosoque for kindly providing additional T. cruzi strains. J. Rivett-Carnac designed the diploid re-sampling software.

Author Contributions

Conceived and designed the experiments: LAM MSL MAM. Performed the experiments: LAM MSL OF TB. Analyzed the data: LAM MSL OF MDL TB MAM JDR. Contributed reagents/materials/analysis tools: OF BA MDL MSL HJC. Wrote the paper: LAM MSL MDL MAM. Conceived the Sylvio X10/I Whole Genome Project: BA OF MAM.

References


3.3.4 Development and evaluation of amplicon deep sequencing markers to characterize intra-host parasite multiclonality

To investigate the extent of natural intra-host parasite multiclonality and its association with congenital infection and clinical status of chronic Chagas disease, two 450 bp multi-copy polymorphic sequence markers (TcGP63 and ND5) were designed, amplified and deep sequenced from 93 clinical primary hemocultures. T. cruzi strains were isolated from chronically-infected patients across the clinical spectrum (asymptomatic to severe cardiomyopathy, megaesophagus or megacolon) and included 46 chronic patients from Goiás, Brazil and 27 chronic patients and 10 mother-infant pairs from Cochabamba, Bolivia.

This study is reported in full below by Llewellyn et al., accepted.

In summary:

- A total of 6,736,749 and 871,855 sequencing reads were generated for ND5 (a mitochondrial gene) and TcGP63 (which encodes a constitutively expressed antigenic surface protease), respectively, and normalized by patient cohort (Goiás – ND5 and TcGP63: 10,000; Cochabamba – ND5: 30,000 and TcGP63: 10,000).
- A series of artificial control mixes of equimolar genomic DNA from DTUs I-IV were also evaluated. Resulting sequence data demonstrated that PCR amplification bias dramatically skewed ST recovery towards the most abundant ST in the mix. A fourfold over representation of a ST in the original mix, resulted in a 100-1000 fold over representation after PCR, which was controlled for in subsequent analyses.
- Three principal ND5 STs were identified among patients, corresponding to TcI, TcII and TcIII-VI. In most clinical samples, STs were monomorphic at the 97% identity level, thus ND5 was used principally to assign populations to DTU-level.
- On the basis of ND5, most Cochabamba chronic cases contained a single ST, likely TcV (25/27), the remaining were TcI (2/27).
- By comparison, sequence diversity was considerably higher in Goiás; TcII was the predominant ND5 DTU identified, and multiple patients were co-infected with minor genotypes at frequencies two orders of magnitude lower (usually TcII/TcIII-TcVI or TcII/TcI).
- Paired congenital cases from Cochabamba resembled chronic cases from the same area in terms of ND5 DTU composition; a subset showed a similar phenomenon to Goiás, with mixed infection profiles (TcI/TcIII-VI) at similar relative abundances (c.1:1000).
- Based on measurements of TcGP63 alpha diversity (Shannon Index), in both Cochabamba and Goiás, there was no clear correlation between intra-host genetic diversity in chronic patients and age, sex or symptoms at the 97% (age p= 0.734 and 0.382; sex p=0.298 and 0.535; symptoms p=0.136 and 0.486, respectively) and 99% (age p=0.854 and 0.319; sex p=0.169 and 0.696; symptoms p=0.00988 and 0.697, respectively) ST divergence level. This might be explained by micro-geographic variation in T. cruzi genetic diversity and/or infection intensity and possible cross-genotype immunity which accumulated with exposure and therefore age.
Between congenital pairs multiple STs were transmitted from mother to infant; in most cases (6/10), similar or greater numbers of *TcGP63* STs were observed in mothers; in four infants novel genotypes were detected. Multiclonal repertoires sampled in mothers and infants will be influenced by parasite tissue sequestration in the mother, stochastic transplacental transfer to the neonate and availability of circulating clones for isolation by hemoculturing.

At the 97% sequence divergence level, when considering well represented *TcGP63* STs, dN/dS ratios among chronic patients were indicative of strong diversifying selection (Goiás, ST1 = 2.6436, ST4 = 6.3415; Cochabamba, ST3=2.8059).

This is the first study to examine parasite multiclonality in individual patients to an unprecedented depth. While this approach demonstrated the resolutive power of amplicon deep sequencing to reveal genetic diversity in chronic and congenital cases, it also highlighted potential biases that can be introduced by the addition of a PCR step and encourages the exploration of novel genome sequencing techniques, including whole genome deep sequencing or single cell genome sequencing, as reference genome assemblies improve and reagent costs decline.
Please be aware that one cover sheet must be completed for each 'Research Paper' included in a thesis.

1. For a 'research paper' already published
   1.1. Where was the work published? .................................................................
   1.2. When was the work published? .................................................................
       1.2.1. If the work was published prior to registration for your research degree, give a brief rationale for its inclusion
       ....................................................................................................................
   1.3. Was the work subject to academic peer review? ........................................
   1.4. Have you retained the copyright for the work? Yes / No
       If yes, please attach evidence of retention.
       If no, or if the work is being included in its published format, please attach evidence of permission from copyright holder (publisher or other author) to include work

2. For a 'research paper' prepared for publication but not yet published
   2.1. Where is the work intended to be published? PLoS Neglected Tropical Diseases
   2.2. Please list the paper’s authors in the intended authorship order
   2.3. Stage of publication – Not yet submitted / Submitted / Undergoing revision from peer reviewers’ comments (In press)

3. For multi-authored work, give full details of your role in the research included in the paper and in the preparation of the paper. (Attach a further sheet if necessary)
The candidate participated in mitochondrial sequencing marker design, provided unpublished reference datasets, laboratory samples, coordinated provision of Bolivian field samples and contributed to data analysis, interpretation and manuscript drafting:

NAME IN FULL (Block Capitals) LOUISA ALEXANDRA MESSENGER
STUDENT ID NO: 223021
CANDIDATE’S SIGNATURE: .......................... Date 10/2/15

SUPERVISOR/SENIOR AUTHOR’S SIGNATURE (3 above) ........................................

Improve health worldwide www.lshtm.ac.uk
Deep sequencing of the *Trypanosoma cruzi* GP63 surface proteases reveals diversity and diversifying selection among chronic and congenital Chagas disease patients.

Llewellyn, MS¹,², Messenger, LA¹, Luquetti AO³, Garcia L⁴, Torrico F⁴, Tavares SBN³, Cheaib, B⁵ Derome, N⁵, Delepine, M⁶, Baulard, C⁵, Deleuze, JF⁶, Sauer, S⁷ Miles, MA¹

¹The London School of Hygiene and Tropical Medicine, London, UK
²Molecular Ecology and Fisheries Genetics Laboratory, School of Biological Sciences, University of Wales, Bangor, Deiniol Road, Bangor, Gwynedd, LL57 2UW, UK
³Laboratório de Pesquisa da doença de Chagas, Hospital das Clínicas da Universidade Federal de Goiás, Brazil
⁴Facultad de Medicine, Universidad Mayor de San Simon, Cochabamba, Bolivia
⁵Institut de Biologie Integrative et de Systemes, Universite de Laval, Quebec, Canada
⁶Centre National de Génotypage, CEA, Evry, Paris, France
⁷Max Planck Institute for Molecular Genetics, Berlin, Germany

Abstract

Background

Chagas disease results from infection with the diploid protozoan parasite *Trypanosoma cruzi*. *T. cruzi* is highly genetically diverse, and multiclonal infections in individual hosts are common but little studied. In this study we explore *T. cruzi* infection multiclonality in the context of age, sex and clinical profile among a cohort of chronic patients, as well as paired congenital cases from Cochabamba, Bolivia and Goiás, Brazil using amplicon deep sequencing technology.

Methodology/ Principal Findings

A 450bp fragment of the trypomastigote TcGP63I surface protease gene was amplified and sequenced across 70 chronic and 22 congenital cases on the Illumina MiSeq platform. In addition a second, mitochondrial target - ND5 - was sequenced across the same cohort of cases. Several million reads were generated and sequencing read depths were normalized within patient cohorts (Goiás chronic, n = 43, Goiás congenital n = 2, Bolivia chronic, n=27; Bolivia congenital, n=20), Among chronic cases, analyses of variance indicated no clear correlation between intra-host sequence diversity and age, sex or symptoms, while principal coordinate analyses showed no clustering by symptoms between patients. Between congenital pairs, we found evidence for the
transmission of multiple sequence types from mother to infant, as well as widespread instances of novel genotypes in infants. Finally, non-synonymous to synonymous (dn:ds) nucleotide substitution ratios among sequences of TcGP63Ia and TcGP63Ib subfamilies within each cohort provided powerful evidence of strong diversifying selection at this locus.

**Conclusions/Significance**

Our results shed light on the diversity of parasite DTUs within each patient, as well as the extent to which parasite strains pass between mother and foetus in congenital cases. Although we were unable to find any evidence that parasite diversity accumulates with age in our study cohorts, putative diversifying selection within members of the TcGP63I gene family suggests a link between genetic diversity within this gene family and survival in the mammalian host.

**Author Summary**

*Trypanosoma cruzi*, the causal agent of Chagas disease in Latin America, infects several million people in some of the most economically deprived regions of Latin America. *T. cruzi* infection is lifelong and has a variable prognosis: some patients never exhibit symptoms while others experience debilitating and fatal complications. Available data suggest that parasite genetic diversity within and among disease foci can be exceedingly high. However, little is know about the frequency of multiple genotype infections in humans, as well as their distribution among different age classes and possible impact on disease outcome. In this study we develop a next generation amplicon deep sequencing approach to profile parasite diversity within chronic Chagas Disease patients from Bolivia and Brazil. We were also able to compare parasite genetic diversity present in eleven congenitally infants with parasite genetic diversity present in their mothers. We did not detect any specific association between the number and diversity of parasite genotypes in each patient with their age, sex or disease status. We were, however, able to detect the transmission of multiple parasite genotypes between mother and foetus. Furthermore, we also detected powerful evidence for natural selection at the antigenic locus we targeted, suggesting a possible interaction with the host immune system.

**Introduction**

*Trypanosoma cruzi* is a kinetoplastid parasite and the causative agent of Chagas disease (CD) in Latin America. *T. cruzi* infects approximately 8 million people throughout its distribution and causes some 13,000 deaths annually [1]. Chagas disease follows a complex course. Infection, often acquired in childhood, is generally lifelong but progression from the indetermined
(asymptomatic) to symptomatic stage occurs in only 30% of cases [2]. A broad pathological spectrum is associated with clinical CD including potentially fatal cardiological and gastrointestinal abnormalities [3]. The relative contributions of parasite and host immunity in driving disease pathology are a matter of continuing debate [4]. Recently, for example, bioluminescent parasite infections in BALB/c mouse models have suggested that heart disease can progress in the absence of detectable local parasite load [5].

It is widely recognized that natural parasitic infections are often comprised of several parasite clones [6]. Malarialogists use the term ‘multiplicity of infection’ (MOI) to describe when multiple Plasmodium sp. genotypes occur within the same host [7,8]. A similar phenomenon has been observed in T. cruzi in vectors (e.g. [9]), as well as mammalian reservoir hosts (e.g. [10]) and humans hosts (e.g. [11]) using solid phase plating and cell sorting techniques. The occurrence of multi-genotype infections has fundamental implications for host immunity [12], as well as for accurate evaluation of pathogen drug resistance [13], transmission rate, epidemiology and population structure (e.g. [7,11]). The efficiency with which it is possible to sample pathogen clonal diversity from biological samples has soared in recent years with the advent of next generation sequencing. Deep sequencing approaches have long been applied to study the dynamics of HIV anti-viral therapy escape mutations. As a result amplicon sequencing increasingly features in a clinical diagnostic context [14]. Plasmodium falciparum MOI can be resolved at merozoite surface protein loci at far greater depths than possible by standard PCR approaches [15]. Furthermore, targeting low copy number antigens in parasite populations via amplicon sequencing can provide important clues to frequency-dependent selection pressures within hosts, between hosts and between host populations [16].

T. cruzi can persist for several decades within an individual host. Unsurprisingly perhaps, therefore, T. cruzi shows significant antigenic complexity. T. cruzi surface proteins are encoded by several large, repetitive gene families that are distributed throughout the parasite genome [17]. Among these gene families the mucins, transialidases, ‘dispersed gene families’ (DGFs), mucin-associated surface proteins (MASPs) and GP63 surface proteases comprise the vast majority of sequences - 10-15% of the total genome size [17,18]. Whilst the role of some of the proteins encoded by surface gene families in host cell recognition and invasion is relatively well understood (e.g. the transialidases [19]), the role of others (e.g. the MASPs, DGFs) is not. Furthermore, the role each plays in evading an effective host response remains largely unknown.

The GP63 surface proteases are found in a wide variety of organisms, including parasitic trypanosomatids [20]. In Leishmania spp. GP63 proteases are the most common component of the parasite cell surface with crucial roles in pathogenicity, innate immune evasion, interaction with the host extracellular matrix and ensuring effective phagocytosis by macrophages [21]. In T.
brucei subsp. The role of GP63 proteins is less well defined, although some protein classes are thought to be involved with variant surface glycoprotein processing between different life cycle stages [22]. In T. cruzi at least four classes of GP63 gene are recognized [20]. Like many GP63 proteases in Leishmania spp., surface expressed T. cruzi GP63 (TcGP63) genes are anchored to the cell membrane via glycosyl phosphatidylinositol moieties [23,24]. Among these are the TcGP63 1a & 1b genes (collectively TcGP63I). TcGP63 1a & 1b encode 78kDa 543 amino acid proteins, are expressed in all life cycle stages and are implicated in the successful invasion of mammalian cells in vitro [23,24].

In the current study we target TcGP63I genes as markers of antigenic diversity among three cohorts of Chagas disease patients: two in Cochabamba, Bolivia and one in Goias, Brazil. We also targeted a maxicircle gene for the NADH dehydrogenase subunit 5 to provide basic T. cruzi genotypic information for each case. Diversity at each of the two T. cruzi loci within each patient was characterized using a deep amplicon sequencing approach, generating several million sequence reads. Our results shed light on the diversity of parasite DTUs within each patient, as well as the extent to which parasite strains pass between mother and foetus in congenital cases.

We were unable to find any evidence that parasite diversity accumulates with age in our study cohorts, or to detect a link between parasite diversity and clinical profile. However, we were able to detect evidence of putative diversifying selection within members of the TcGP63 gene family, suggesting a link between genetic diversity within this gene family and survival in the mammalian host.

Materials and Methods

Ethical Statement Ethical permissions were in place at the two centres where human sample collections were made, as well as at the London School of Hygiene and Tropical Medicine (LSHTM). Local ethical approval for the project was given at the Plataforma de Chagas, Facultad de Medicina, UMSS, Cochabamba, Bolivia by the Comite de Bioetica, Facultad de Medicina, UMSS. Local ethical permission for the project was given at the Hospital das Clinicas da Universidade Federal de Goias (UFG), Goias, Brazil by the Comite de Etica em Pesquisa Médica Humana e Animal, protocol number 5659. Ethical approval for sample collection at the LSHTM was given for the overall study, “Comparative epidemiology of genetic lineages of Trypanosoma cruzi” protocol number 5483. Samples were collected with written informed consent from the patient and/or their legal guardian.

Biological sample collection: Parasite isolation protocols were different between centres. At the UMSS, 0.5 mL of whole venous blood was taken from chronic patients and inoculated directly into biphasic blood agar culture. T. cruzi positive samples were minimally repassaged and
cryopreserved at log phase (precise repassage history unavailable). For infants, 0.5 mL of chord blood was taken at birth and inoculated into culture. Again, positive samples were cryopreserved at log phase after minimal repassage (precise repassage history unavailable). DNA extractions, using a Roche High-Pure Template Kit, were made directly from the cryopreserved stabilate. At the UFG, 17 mL of whole blood was collected into EDTA, centrifuged for 10 minutes at 1200g at 4 °C and the plasma replaced with 8mL Liver Infusion Tryptone (LIT) medium. After a further 10 minutes at 1200g (4°C), the supernatant was again removed. Two mL of packed red blood cells were subsequently transferred to 3 mL of LIT medium and checked periodically for signs of epimastigote growth by light microscopy. Positive cultures were not repassaged. Instead primary cultures were stabilized by the addition of guanidine 6 M-EDTA 0.2 M (Sigma-Aldrich, UK). DNA extractions were made from the full volume using the QIAamp® DNA Blood Maxi Kit (Qiagen, UK) according to the manufacturer’s instructions. Among Bolivian strains, DNA concentrations submitted to PCR were standardized after quantitation using a PicoGreen® assay. In view of presence of human genetic material in Goias samples, parasite DNA concentrations were standardized to within the same order of magnitude via qPCR as previously described [25]. All samples collected for in this study are listed in Table 1.

Epidemiological and clinical observations: The two areas studied have dissimilar histories in terms of Chagas disease transmission intensity. Vector-borne T. cruzi transmission in Goias and its surrounding states (where samples were collected – Table 1) was interrupted approximately 20 years before the sampling detailed in this study [26,27]. In the sub-Andean semi-arid valleys of Cochabamba and its environs, however, vector-borne domestic transmission is still a likely source of new infections, albeit at a reduced rate since intensive spraying campaigns in the mid 2000s [28]. Clinical data collected in this study were categorised simply into symptomatic and asymptomatic classes for statistical tests in view of samples sizes. Sub-categories within symptoms were defined as 1) Cardiopathy (including any electrocardiographic and/ or echocardiographic abnormalities, X-ray with cardiac enlargement. Patients with atypical cardiac abnormalities i.e. those not exclusively associated with Chagas disease, were included in the symptomatic class in the context of this study.) 2) Megaesophagus (including achalasia and barium swallow abnormalities) 3) Megacolon (constipation associated with dilation as by barium enema) and 4) Normal (no symptoms or signs on examination and a normal electrocardiogram) (Table 1)

Primer design, PCR conditions, amplicon sequencing and controls: Degenerate primers for a 450bp fragment of the maxi-circle NADH dehydrogenase 5 were designed as described in Messenger et al. 2012 [29]. Degenerate primer design for the TeGP63I family surface proteases
(including Ia and Ib subclasses) [24] was achieved by reference to sequences retrieved from EuPathDB for Esmeraldo (TcII), CL Brener (TcVI), Silvio (TcI) and JR (TcI) (http://eupathdb.org/). Primer binding site positions in relation to TcGP63I putative functional domains are displayed in Figure S1. Homologs were identified by BLAST similarity to a complete TcGP63I sequence (bit score (S) ≥ 1000). Alignments of resulting sequences were made in MUSCLE [30] and primers were designed manually to target a variable region within and between individual strains with a final size of 450bp. ND5b primer sequences were ND5b_F ARAGTACACAGTGGRYYTRCAYA; ND5b_R CTTGCYAARATACAACCACAA. The final TcGP63 primers were TcGP63_F RGAACCGATGTCATGGGGCAA and TcGP63_R CCAGYTGGTGTAATRCTGCYGCC. Amplification was undertaken using the Fluidigm® platform and a reduction of the manufacturer’s recommended number of cycles to a total of 26 was made in an attempt to minimise PCR amplification bias. Thus, the manufacturer’s recommended conditions were adapted to the following protocol: one cycle of 50°C for 2 minutes, 70°C for 20 minutes, and 95°C for 10 minutes; six cycles of 95°C for 15 seconds, 60°C for 30 seconds, 72°C for 60 seconds; two cycles of 95°C for 15 seconds, 80°C for 30 seconds, 60°C for 30 seconds and 72°C for 60s; five cycles of 95°C for 15 seconds, 60°C for 30 seconds, 72°C for 60 seconds; two cycles of 95°C for 15 seconds, 80°C for 30 seconds, 60°C for 30 seconds and 72°C for 60 seconds; five cycles of 95°C for 15 seconds, 60°C for 30 seconds, 72°C for 60 seconds, and finally five cycles of 95°C for 15 seconds, 80°C for 30 seconds, 60°C for 30 seconds and 72°C for 60 seconds. Amplifications were performed using the FastStart High Fidelity PCR System (Roche). Three PCR reactions were pooled per sample prior to sequencing in an attempt to further reduce amplification biases [31]. Equimolar concentrations of ND5 and TcGP63I amplicons from 96 DNA samples were multiplexed on Illumina runs using dual index sequence tags (Illumina Inc). Sequencing was undertaken using a MiSeq platform using a 2 x 250 bp (Reagent Kit version 2) according to the manufacturer’s protocol. In addition to the clinical samples, we included a dilution series of control samples. The controls comprised artificially mixes of DTUs I-VI genomic DNA at equimolar concentrations. At the ND5 locus, comparison between the expected DTU abundance ratios and diversity of artificial control mixes and that defined via amplicon sequencing was made (Figure S2).

**Amplicon sequence data analysis** – De-multiplexed paired-end sequences were submitted to quality control and trimming in Sickle [32] and mate pairs trimmed in FASTX Toolkit (http://hannonlab.cshl.edu/fastx_toolkit/). ND5, TcGP63 and contaminating sequences were then sorted against a reference using BOWTIE2 [33]. Individual paired reads were found to be overlapping in only a minority of cases. Thus we chose to proceed with analysis of a sequence fragment with a truncated central section for both targets. Further sequence manipulations were undertaken using FASTX Toolkit and custom awk scripts to parse files and concatenate each mate
pair into a single sequence for downstream analysis. MUSCLE [30] was used for alignment of amplicon sequences in each patient sample. Next, analysis was undertaken in the Mothur software package [34] for the elimination of putative PCR chimeras and individual sequence clustering. The Shannon index of diversity was calculated at the intra-patient level based on sequence types (STs) defined at 97% and 99% identity in Mothur [34]. Comparisons of Shannon diversity were made between patients in each cohort (Bolivia chronic, Bolivia congenital, Goias chronic) via analyses of covariance and linear regression in the R package (http://CRAN.R-project.org).

TcGP63I sequence datasets for patients from each cohort were then merged and analyses conducted using 97% and 99% STs defined with UPARSE [35] across patients. Weighted UniFrac distances between TcGP63I STs among samples were generated and subsequently clustered via a principal coordinates analysis in QIIME [36]. Estimates of diversifying selection among TcGP63I STs were made in KaKs Calculator [37] using Yang and Neilson’s 2000 approximate method [38] and tested for significance using a Fisher’s exact test. Prior to selection calculations, sequences were clustered into 99% identity STs and singletons excluded in an attempt to exclude SNPs introduced as PCR artefacts. To test for diversifying selection across putative TcGP63I gene families (TcGP63Ia & Ib - 97% cut-off as defined by Cuevas and colleagues [24]), 99% identity STs from each patient cohort were pooled (Table 2). To test for selection within TcGP63I gene families, STs within each 97% category (corresponding to TcGP63Ia & b respectively) were examined separately per cohort (Table 2).

All amplicon sequences generated in this study are freely available from the authors on request.

**Results**

**Sequence yields and discrete typing unit (DTU) designations.** After quality filtering, trimming, decontamination and removal of unpaired reads, 6,736,749 reads were assigned to the ND5 mitochondrial marker and 871,855 to TcGP63I marker across the 92 clinical samples, perhaps reflecting higher copy number in the former than the latter. After trimming, the overlap between individual mate pairs was marginally too short to be assembled into a single read. Thus paired reads were first aligned against a full-length reference fragment and the central portion excised to remove any gaps and ensure correct alignment. Sequence depth thresholds per sample for inclusion were set for each dataset (Goias - ND5 & TcGP63 – 10,000; Cochabamba – ND5: 30,000; TcGP63 10,000; see Figure 1). Reads from samples in excess of this threshold were discarded and samples with read counts below this threshold excluded. Our aim in setting the threshold was: 1) To include as many samples as possible while maintaining a good depth of coverage; 2) To standardise sampling intensity across individuals and thus facilitate comparisons between them.

The ND5 mitochondrial target was sequenced to provide DTU I-VI identification of parasites circulating within and among patients by comparison to existing data [29]. However, with
reference to the results from the control samples - and due the necessary truncation of the
sequence fragment - only three groups could be reliably distinguished, corresponding to the three
major *T. cruzi* maxicircle sequence classes [39]. The three groups corresponded to TcI, TcII and
TcIII-VI respectively. Furthermore, in reference to the control mixes, we found evidence that
amplification bias dramatically skewed the recovery of sequence types (STs) towards the TcIII-VI
group. Some skew is expected, as these four DTUs (TcIII-VI) share the same maxicircle sequence
class, and this class is thus more abundant in the control mix. However, TcI and TcII - which
should have in theory been present as 16% (1/6) of all sequences in the controls respectively -
were in fact present (on average) at only 2.9% and 0.03% among the four samples where all three
STs were recovered (Figure S2). Amplicon sequencing from the two most concentrated controls
(57 ng/μL and 125 ng/μL genomic DNA respectively) resulted in poor sequence yields and a
failure to recover all three STs.

Unsurprisingly perhaps in the light of the control data, most clinical samples were dominated by
sequences from a single group, with minor contributions from others (Figure 2). Indeed sequences
recovered from many strains were monomorphic at the 97% identity level - especially in
Cochabamba. As such, comparisons based on ND5 are necessarily descriptive and meaningful
alpha (within sample) and beta (between sample) diversity statistics were not calculated. Figure 2
shows the distribution of DTUs among samples as defined by the ND5 locus. Most Cochabamba
chronic cases samples were assigned to a single sequence within the TcIII-VI group (likely to be
TcV, as we defined with standard genotyping assays [40] with the exception to two TcI cases –
PCC 240 and PCC 289 (Figure 2, Panel B). Sequence-type diversity in Goias was considerably
higher (Figure 2, Panel A). In this case the TcII group, rather than the TcIII-VI group,
predominated. Unlike in Bolivia, sequences from other groups were present alongside TcII in
multiple patients but at frequencies two orders of magnitude lower. Congenital pairs that
originated from Cochabamba resembled chronic cases from the same region in their DTU
composition (TcIII-VI group predominant, Figure 2, Panel C). Strikingly, mother/child pair
CIUF65 (B5) and CIUF75 (M5) share similar mixed infection profiles (TcI/ TcIII-VI) at similar
relative abundances (c.1:1000), consistent with the minor to major genotype abundance ratios
observed in Goias. The same is also true for the Goias congenital pair (Figure 1) which both
showed TcII/TcI mixes. Finally, sequential isolates taken from the same Goias chronic patient at
different time points suggest that minor abundance genotypes are not always consistently
detectable in the blood (Figure 2): TcI is absent at first sampling of patient y, but present at the
second sampling. For patient z, the TcIII-VI genotype is only present in the first of the two sample
points. For both Cochabamba and Goias, reference to the control data suggests that ‘minor
genotypes’ could be substantially more abundant in the patients than the amplicon sequence data
suggest.
**TcGP63I surface protease alpha diversity among clinical and congenital cases**

Alpha diversity measurements aim to summarise the diversity of species (in this case STs), within an ecological unit (in this case a host). We summarized the number of STs and their relative abundance in each of our samples, using the Shannon Index (SI) [41]. Among non-congenital cases, our aim was to evaluate possible associations between TcGP63I antigenic diversity and several epidemiological and clinical parameters - age, sex and disease status. We used analyses of covariance (ANCOVA) to test for the effect of these parameters on intra-host antigenic diversity (STs defined both at 97% and 99% for comparison), combining continuous (age) and categorical (sex, clinical forms) data. In Cochabamba, regardless of the order in which parameters were included as factors in the model, there was no evidence for a main effect of age, sex or symptoms on alpha diversity (SI) at either ST divergence level (97% ST Age: p = 0.734; Sex: p = 0.298; clinical form: p = 0.136. 99% ST - Age: p = 0.854; Sex: p = 0.169; clinical form: p = 0.0988).

Similarly, ANCOVAs were non-significant for an association between the SI and age, sex or symptoms in Goias (97% ST - Age: p = 0.382; Sex: p = 0.535; clinical form: p = 0.486. 99% ST - Age: p = 0.319; Sex: p = 0.696; clinical form: p = 0.697). Finally, we undertook linear regressions of SI with age in each population. As one might expect from previous ANCOVAs, no significant correlation was detected (Goias $R^2= 0.0233, p=0.340$ (97% ST); $R^2= 0.0256, p=0.3049$ (99% ST) Cochabamba $R^2=0.0287, p=0.429$ (97% ST); $R^2= 0.0230 p = 0.479$(99% ST)).

Congenital comparisons were made pairwise between mother and infant at 99% ST similarity. In addition to the ten matched isolate pairs from Cochabamba, a single pair from Goias was also included (6718 & 6720) in the comparisons. The results of the alpha diversity comparisons are shown in Figure 3, and read depths were balanced between samples. In terms of the absolute number of STs identified, infants exceeded mothers in most instances (pairs 2, 3, 4, 5, 6, 8 & 9).

In the remaining cases however (4/11), the number of antigenic sequence types was greater in the mother. Shannon diversity index comparisons between mothers and infants, which also takes ST abundance into account, suggested that some differences (e.g. pairs 4, 5 &6) might be marginal (Figure 3).

**TcGP63I ST distributions among clinical and congenital CD patients.** Individual sample sequence datasets within each of the different study cohorts (Cochabamba congenital, Cochabamba non-congenital and Goias) were merged to facilitate analysis of the distribution of antigen 99% STs among individuals (i.e. beta-diversity comparisons). Pairwise weighted Unifrac distances were calculated within cohorts of chronic cases from Cochabamba and Goias to examine whether the sequence diversity of the TcGP63I antigenic repertoire present in each patient could be associated with disease outcome. Principal coordinate analyses of the resulting matrices are
displayed in Figure 4. Among cases from Goias, repertoires varied considerably among cases, with several outliers. However, repertoires from symptomatic and asymptomatic cases were broadly overlapping in terms of sequence identity, and no clustering was noted among different symptom classes either (Figure 4, Plot B). TcGP63I read yields permitted comparisons for only two pairs of sequential isolates from the sample patients – $x$ and $y$ (see Table 1) – both of which showed closely clustering, although non-identical, profiles. TcGP63I diversity between Cochabamba chronic cases was arguably lower, with the exception of two outliers unambiguously identified as TcI with reference to the ND5 locus (all others were classified as TcIII-VI - likely TcV). Again, however, symptomatic and asymptomatic cases were broadly overlapping.

Sequence type profile comparisons among Cochabamba congenital cases were made for 99% STs and are displayed in heatmap format in Figure 5. There are two key features of interest. The first is that profiles in mother and infant can match very closely (e.g. pairs 2&6). The second is that novel STs were present in the infant sample with respect to the mother in half of the cases. Indeed, in pair 9, the infant profile was radically different to that of the mother.

**Population-level Ka/Ks ratios within and between TcGP63I gene family members.** Trimmed TcGP63 reads, pre-filtered for quality and PCR errors, were pooled within each study site (Bolivia, Goias). To further reduce minority SNPs and PCR errors, STs were defined at 99% with each site in UPARSE [35]. $Ka/Ks$ ratio estimates within each study area indicated a significant excess of synonymous mutations among STs (Goias = 0.8354, Bolivia = 0.7515) averaged across sites (Table 2). However, when calculations were based on diversity present among well represented STs of each gene family member (TcGP63Ia and TcGP63Ib, 97% cut-off [24]) a powerful and significant excess of non-synonymous substitutions was noted within each study area ($Ka/Ks$, Goias, ST1 = 2.6436, ST4 = 6.3415; Bolivia ST3 = 2.8059; Table 2). Again, calculations were based not on individual sequences, but rather 99% STs within predefined 97% clusters. The position of the 97% STs in question is shown in the tree in Figure S3, with clear similarity between those clusters under apparent diversifying selection (Goias ST1 & 2, and Bolivia ST3) with TcGP63Ia and TcGP63Ib references respectively [24].

**Discussion**

In this study our aim was to collect a cohort of *T. cruzi* samples from clinical CD cases, representative of different endemic regions and of different ages and disease presentations, to explore links between CD epidemiology and multiplicity of infection. To provide a robust, sensitive and quantifiable means of assessing intra-host parasite diversity we first implemented standardized parasite isolation (and enrichment) strategies within each study cohort. Latterly, we
developed an amplicon sequencing approach to profile parasite diversity within each patient. Given the relatively short (400-500bp) read lengths generated by next generation sequencing platforms (at the time of experimentation), we chose a rapidly evolving maxicircle gene (ND5) in an attempt to resolve DTU level diversity ([29]). Current multilocus nuclear targets are generally too long (500bp+) to meet our selection criteria [42]). To explore antigenic diversity, we chose a putatively low (5-10) copy number gene family member TcGP63I, expressed on the parasite surface during the amastigote and trypomastigote lifecycle stage and thus exposed to the human immune system [24]. Given that both ND5 and TcGP63I are present as several copies per parasite genome (and potentially show inter-strain copy number variation e.g. [43]), one cannot presume a 1:1 relationship between ST and parasite individual, even if we were able to account for the PCR amplification bias we detected. The identification of a genetically, variable, single copy, surface expressed antigen locus is a major challenge in T. cruzi – antigen genes are by their nature highly repetitive [17,18]. TcGP63I, with its relatively low copy number represents the closest currently available fit, and, as we have shown, provides a useful target for revealing intra-host antigenic diversity. Merozoite surface proteins (MSP) 1 and 2 have traditionally provided useful targets for detecting MOI in P. falciparum (e.g. [44,45]). Furthermore, amplicon sequencing of the MSP locus has been successfully proven to reveal as many as six-fold more variants than traditional PCR-based approaches [15].

The substantial historical interest in defining MOI among P. falciparum owes itself to the strong correlation between MOI and rate of parasite transmission [46]. As such, fluctuations in transmission intensity can be tracked to evaluate the efficiency of vector eradication campaigns, drug treatments, the introduction of insecticide-treated nets etc – without the need to directly estimate the entomological inoculation rate. Evaluation of CD transmission intensity has its own challenges. The presence of infected individuals, triatomine vectors in domestic buildings, incrimination of vectors via human blood meal identification (e.g. [47]) can all help to build the overall picture. However, parasite transmission is likely to occur in only a tiny proportion of blood meals [48,49], and vector efficiency is thought to vary considerably between triatomine species [50] - thus the presence of vectors is no guarantee of transmission. Infection with T. cruzi is lifelong, thus positive patient serology is not a reliable indicator of active parasite transmission either. Traditionally, active T. cruzi transmission has been implied from positive serology among younger age classes. Especially in hyperendemic areas of Bolivia, Paraguay and Argentina the proportion of seroprevalent individuals increases with age [51,52]. MOI in T. cruzi patients should follow a similar trend given a stable force of infection. Furthermore MOI comparisons between disease foci could, controlling for age, facilitate an appreciation of relative transmission intensities – a useful tool for those who wish to track the efficacy of interventions. In the current study, however, we were unable to identify a correlation between MOI and age, even once patient sex
and clinical form had been corrected for. Our inability to validate this fundamental prediction has many possible causes. First, patients in each cohort originate from different communities within each study area (Table 1). Micro-geographic variation in *T. cruzi* genetic diversity is commonly observed (e.g. [11,53,54], and the same is likely to be true for infection intensity. Thus, if patients from different sites share dissimilar histories in the intensity and diversity of exposure to *T. cruzi* clones, comparisons between them are difficult to make. Secondly, the relationship between MOI and age is not necessarily linear. If a degree of cross-genotype immunity accumulates with exposure, one might expect a slower increase in intra-host antigenic diversity in older age groups. However, this was not the case in our dataset and neither a linear, nor a unimodal relationship could be established.

Amplicon sequencing approaches to the study of transmission patterns in human parasites have so far been restricted to those species that replicate and reach high parasitemias in peripheral blood (i.e. *T. brucei* [55] and *P. falciparum* [13,15]). *T. cruzi* trypomastigote circulating parasitemias, as measured by qPCR, are thought to vary considerably between acute (400 parasites/ml), newborn (150-12000 parasites/ml) and chronic (3-16 parasites/ml) cases [25,56]. Nonetheless, they remain several orders of magnitude lower than those that occur during *T. brucei* or *P. falciparum* infections. Low circulating *T. cruzi* parasitemia presents major problems to studies that aim to achieve molecular diagnosis of CD in chronic cases and ours is no exception. One problem is that much of the parasite diversity present in the host is likely to be sequestered in the tissues at any give time [57], as our sequential samples from Goias also suggest. Thus blood stage parasite genetic diversity may be a poor representation of that actually present in the host. Another confounder is culture bias, by which differential growth of clones in culture, as well as loss of clonal diversity during repassage can both influence diversity estimates. Attempts to generate amplicon sequence data directly from clinical blood samples would likely to be thwarted by low circulating parasitemia [25, 56]. Instead we elected to enrich for parasite DNA via culture – in Goias without further repassage, but in Bolivia with at least one repassage before cryopreservation.

Low circulating parasitemia in Chagas patients also means it is possible that amplicon-sequencing strategies might rapidly ‘bottom out,’ if few parasites are present within a sample. In our dataset, for example, at the ND5 locus, minority DTUs at 97% divergence can be present as a proportion of < 1 in 1000 (Figure 1), with the implication that several thousand parasites must be present in the sample. In both Goias and Bolivia matched instances occurred in congenital cases where TcI exists in mother and infant as the minor DTU at similar relative abundance (i.e. 1 in 1000, Figure 1). It is highly unlikely that these data directly reflect chronic CD parasitemia levels. Instead, with reference to the data we obtained from the controls, PCR amplification bias is a more likely source of unrealistic major to minor genotype ratios. As such, the fourfold over-representation of a ST in the original sample, for example, can result in 100-1000 fold over-representation after PCR.
However, while the relative abundance of sequence types recovered using the amplicon approach may be an inaccurate reflection of those present for both ND5 and TcGP63, similar profiles between mother and infant suggests that this bias is likely to be consistent across samples. Thus comparisons between samples are still valid. Furthermore for ND5 at least it seems that *T. cruzi* frequently exchanges mitochondrial (maxicircle) genomes with little apparent evidence of nuclear exchange [11,29]. Fusion of maxicircle genomes occurs transiently during *T. brucei* genetic exchange events [58], and may also do so in *T. cruzi*. Even though standard maxicircle genotyping of progeny only ever reveals a single parent in both species, it is possible that heterologous maxicircle sequences may persist at low abundance in parasite clones. Such a phenomenon could explain the DTU sequence type ratios observed, and this study is the first to sequence a maxicircle gene to this depth.

There is general consensus in the literature is that the likelihood of congenital CD transmission is not strongly influenced by the genotype of the parasite infecting the mother [59-61]. Nonetheless, the majority of cases are reported in the Southern Cone region of South America, providing a circumstantial link with major human-associated *T. cruzi* genotypes TcV TcII, and TcVI. In this study, in the one mixed infection we found, major and minor DTUs (TcVI / TcI) detected in the mother at the ND5 locus were recovered from the infant in similar proportions. TcGP63I beta diversity comparisons of STs defined at 99% showed substantial sharing of between mother and infant (Figure 5). However, both beta diversity comparisons (Figure 5) and total ST diversity (alpha) comparisons (Figure 3) at 99% indicate that while maternal diversity sometimes exceeds that of the infant (explicable perhaps by sequestration in the mother and selective or stochastic trans-placental transfer), the reverse is frequently true. The occurrence of STs in the infant, not present in the mother, has several possible explanations. The infants sampled in this study were neonates, thus superinfection can be ruled out as a source of further parasite clonal diversity. A recent study of infected neonates in Argentina estimated mean infant parasitemia at 1,789 parasites/ml via qPCR - far in excess of that one might expect in the mother [56]. Thus the parasite sample size discrepancy between mother and infant perhaps explains the unexpected levels of diversity in the infant. Even though the TcGP63I gene family is apparently under intense diversifying selection, it seems unlikely that point mutation could generate novel variants over such a short time scale to explain genetic diversity in the infant. Structural variants and homologous recombination are a potential source of diversity, although most, if not all of recombinants should have been excluded in the quality filtering stages, and would be hard to distinguish from PCR chimeras in any case.

Many important *T. cruzi* surface genes belong to large, recently expanded paralogous multigene families [17]. The abundance of these gene copies highlights their likely adaptive significance in
terms of infectivity and host immune evasion, especially because trypanosomatids exert so little control of gene expression at the level of transcription [62]. In *Leishmania major*, for example, it has been recently shown that gene amplification may rapidly duplicate segments of the genome in response to environmental stress [63]. As well as expansion, adaptive change is also likely to occur at the amino acid level among members of paralogous gene families, as has been suggested for *T. brucei* [64]. Despite the relatively small size of the TcGP63I gene family, the amplicon sequencing approach we employed allowed us to explore selection at the level of the gene within the population, i.e. within and between parasite genomes within and between hosts at the population level. Highly elevated non-synonymous substitutions suggest intense diversifying selection within TcGP63Ia and TcGP63Ia STs respectively for those assigned to TcII or TcI. STs from patients infected with TcIII-TcVI (putative TcV) showed few apparent substitutions (Table 2), perhaps consistent with the recent origin of this DTU [65]. The sequence fragment we studied was outside the zinc binding domain of this metalloprotease, indicating selective forces can act on this protein independent of its core proteolytic function, perhaps through repeated exposure to host immunity.

It is important not to overlook the potential importance of multiclonal infections for parasitic disease, both as markers of population level factors such as parasite transmission, but also at the host level, including immunity and disease progression. In this study we have developed an amplicon sequencing approach to probe parasite genetic diversity within and among clinical CD cases to unprecedented depth. While our approach shows the power of this amplicon-seq to resolve diversity in clinical and congenital CD cases, it also highlights the potential biases that might be introduced with the addition of a PCR step. A tool that allows the accurate evaluation MOI would be valuable for tracking transmission rates at restricted disease foci (i.e. villages, outbreaks) in the context of measuring the success of intervention strategies. A similar tool could provide a powerful means of longitudinal tracking of *T. cruzi* infections in terms of disease progression, treatment failure and immunosuppression. Here we demonstrate that amplicon sequencing could have a role to play in this context. However, as sequencing costs decline and reference genome assemblies improve, whole genome deep sequencing, perhaps even of individual parasite cells, becomes and increasingly viable option as it already has for *Plasmodium sp.* [7,66].

**Acknowledgements.** We gratefully acknowledge the assistance of A Boland and R Olaso at the CNG for the help with sample processing. M Lewis at the LSHTM provided valuable comments in the analytical stages. S. Creer at MEFGL Bangor provided helpful suggestions for appropriate software tools and analyses.
References


**Figures and Tables**

**Figure 1 – Read depths by sample and locus for Goias and Cochabamba chronic patient cohorts after quality filtering.** Read depths generated on the Illumina MiSeq platform were standardized across samples prior to analysis. Inclusion thresholds for TcGP63 (Goias – 10,000;
Cochabamba – 3000; wide dash line; red bars) and ND5 (Goais - 10,000; Cochabamba - 30,000; thin dash line; blue bars) are shown for each population.

Figure 2 – Bar plot showing sequence type identity and abundance defined at 97% similarity for the ND5 locus across all samples. A – Goias cohort chronic/intermediate cases; B - Cochabamba chronic/intermediate cases; C – Cochabamba congenital cases. Y axes show log transformed abundance (read counts). X axes show clustered bars for individual samples. Sequence type identities are given in the legend. Stars denote congenital pair from Goias. Labels x (6416 / 6452), y (6401 / 6536) and z (6379 / 6445) sample pairs from the same patient at different time points (see Table 1).

Figure 3 – Alpha diversity indices for TcGP63I amplicon diversity derived from pairs of congenital Chagas disease cases. Diversity indices were derived from STs defined at 99% sequence similarity. Bar plot and associated x-axis on the right hand side shows the Shannon diversity index calculated in Mothur [34], with error bars defining upper and lower 95% confidence intervals.

Figure 4 – Principal coordinates analysis of sequence diversity between chronic Chagas Disease patient TcGP63I antigenic repertoires. Genetic distances are based on a weighted unifrac metric. Plot A shows diversity comparisons among Goias asymptomatic (asympt) and symptomatic (sympt) clinical cases, as well as one acute case. Plot B shows Goias cases with symptoms categorised as acute, card (cardiopathy), card + mega (cardiopathy as well as megacolon and / or megaesophagus), mega (megacolon and / or megaesophagus) or asympt (asymptomatic). Plot C shows comparisons among Cochabamba clinical cases (not including congenital cases) classified as either asymptomatic (asympt) and symptomatic (sympt). The dashed circle on plot C indicates samples unambiguously defined as TcI at the ND5 locus. Pairs of sequential isolates from the same patient are labelled x and y respectively.

Figure 5 – Heatmap comparing the TcGP63I antigenic repertoires of mother and infant congenital pairs. Pairs are indicated down the left hand side of the image (y axis). The mid-point rooted maximum likelihood tree on the x axis describes relationships among the 99% similarity sequence types (STs) identified in UPRASE [35] and was generated in Topali under equal-frequency transversion model, allowing gamma distributed weights across sites [67]. Values on dendrogram notes indicate % bootstrap support. Starred congenital pairs are those where STs are present in the infant but not in the mother.

Figure S1 – TcGP63Ia and Ib amino acid alignments showing amplicon seq primer binding sites in relation to putative functional domains. Amino acid sequences are derived for those
define by Cuevas and colleagues [24]. The colour key on the left hand side indicates primer
binding sites and functional domains. The green shaded regions indicate the area covered by the
Illumina paired end reads along each amplicon. The purple shaded central region indicates the
area not covered.

**Figure S2** - Bar plot of amplicon sequence data generated from control DTU mixes.
Expected ratios of ND5 sequence types (far right) are compared to those recovered via amplicon
sequencing. All three sequence types (I, II, III-VI) were recovered from all but the two most
concentrated control mixes. However, the relative proportions of each sequence type derived
from amplicon sequence data were radically different to that expected.

**Figure S3** – Maximum likelihood phylogeny of 97% TcGP63I STs derived in this study and
available *T. cruzi* and *T. cruzi marinkellei* TcGP63 paralogues. Homologous sequences were
recovered from www.TriTrypDB.org via BLAST. The appropriate substitution model was defined
as the transversion model with invariable sites plus gamma in Topali [67]. Abundant ST labels
correspond with those indicated in Table 2. Branches are coloured by source DTU or red, for
sequences generated in this study. Reference sequences TcGP3Ia and TcGP63Ib from the
literature are also shown alongside 97% sequence types generated in this study [24].
Table 1 – Samples provenance and symptoms

<table>
<thead>
<tr>
<th>Code</th>
<th>ND</th>
<th>Type</th>
<th>NDSequence</th>
<th>Symptoms</th>
<th>Province</th>
<th>Country</th>
<th>Shannon Index TcGp63</th>
<th>Shannon Index TcIII</th>
</tr>
</thead>
<tbody>
<tr>
<td>PCC 226</td>
<td>2.26465</td>
<td>ND</td>
<td>ND</td>
<td>ND</td>
<td>ND</td>
<td>ND</td>
<td>ND</td>
<td>ND</td>
</tr>
<tr>
<td>PCC 326</td>
<td>2.32889</td>
<td>ND</td>
<td>ND</td>
<td>ND</td>
<td>ND</td>
<td>ND</td>
<td>ND</td>
<td>ND</td>
</tr>
<tr>
<td>PCC 426</td>
<td>2.42978</td>
<td>ND</td>
<td>ND</td>
<td>ND</td>
<td>ND</td>
<td>ND</td>
<td>ND</td>
<td>ND</td>
</tr>
<tr>
<td>PCC 526</td>
<td>2.53087</td>
<td>ND</td>
<td>ND</td>
<td>ND</td>
<td>ND</td>
<td>ND</td>
<td>ND</td>
<td>ND</td>
</tr>
<tr>
<td>PCC 626</td>
<td>2.64096</td>
<td>ND</td>
<td>ND</td>
<td>ND</td>
<td>ND</td>
<td>ND</td>
<td>ND</td>
<td>ND</td>
</tr>
<tr>
<td>PCC 726</td>
<td>2.75105</td>
<td>ND</td>
<td>ND</td>
<td>ND</td>
<td>ND</td>
<td>ND</td>
<td>ND</td>
<td>ND</td>
</tr>
<tr>
<td>PCC 826</td>
<td>2.86114</td>
<td>ND</td>
<td>ND</td>
<td>ND</td>
<td>ND</td>
<td>ND</td>
<td>ND</td>
<td>ND</td>
</tr>
<tr>
<td>PCC 926</td>
<td>2.97123</td>
<td>ND</td>
<td>ND</td>
<td>ND</td>
<td>ND</td>
<td>ND</td>
<td>ND</td>
<td>ND</td>
</tr>
<tr>
<td>ID</td>
<td>Sex</td>
<td>Age</td>
<td>Location</td>
<td>Disease</td>
<td>Tc Stage</td>
<td>Country</td>
<td>City</td>
<td>Province</td>
</tr>
<tr>
<td>--------</td>
<td>-----</td>
<td>-----</td>
<td>-------------------</td>
<td>------------------</td>
<td>----------</td>
<td>---------</td>
<td>------------------</td>
<td>----------</td>
</tr>
<tr>
<td>6386</td>
<td>F</td>
<td>39</td>
<td>Rubiataba, Goias</td>
<td>Megaesophagus</td>
<td>TcII</td>
<td>Brazil</td>
<td>Brazil</td>
<td>Goias</td>
</tr>
<tr>
<td>6387</td>
<td>M</td>
<td>30</td>
<td>Angical, Bahia</td>
<td>Asymptomatic</td>
<td>TcII</td>
<td>Brazil</td>
<td>Brazil</td>
<td>Bahia</td>
</tr>
<tr>
<td>6388</td>
<td>M</td>
<td>30</td>
<td>Jussara, Goias</td>
<td>Megaesophagus</td>
<td>TcII</td>
<td>Brazil</td>
<td>Brazil</td>
<td>Goias</td>
</tr>
<tr>
<td>6389</td>
<td>F</td>
<td>37</td>
<td>Quilombo, Comitandia, Goias</td>
<td>Asymptomatic, Cardiopathy, Megacolon, Megaesophagus</td>
<td>TcII/TcI</td>
<td>Brazil</td>
<td>Brazil</td>
<td>Goias</td>
</tr>
<tr>
<td>6390</td>
<td>M</td>
<td>32</td>
<td>Cocos, Bahia</td>
<td>Cardiopathy</td>
<td>TcII</td>
<td>Brazil</td>
<td>Brazil</td>
<td>Bahia</td>
</tr>
<tr>
<td>6400</td>
<td>F</td>
<td>47</td>
<td>Correntina, Bahia</td>
<td>Asymptomatic</td>
<td>TcII/TcI</td>
<td>Brazil</td>
<td>Brazil</td>
<td>Bahia</td>
</tr>
<tr>
<td>6401</td>
<td>M</td>
<td>30</td>
<td>SantaMariaVitoria, Bahia</td>
<td>Cardiopathy</td>
<td>TcII</td>
<td>Brazil</td>
<td>Brazil</td>
<td>Bahia</td>
</tr>
<tr>
<td>6403</td>
<td>F</td>
<td>44</td>
<td>Correntina, Bahia</td>
<td>Cardiopathy</td>
<td>TcII</td>
<td>Brazil</td>
<td>Brazil</td>
<td>Bahia</td>
</tr>
<tr>
<td>6404</td>
<td>F</td>
<td>37</td>
<td>Sao Luiz MBelos, Goias</td>
<td>Asymptomatic</td>
<td>TcII</td>
<td>Brazil</td>
<td>Brazil</td>
<td>Goias</td>
</tr>
<tr>
<td>6405</td>
<td>F</td>
<td>37</td>
<td>Sacaba, Cochabamba</td>
<td>Asymptomatic</td>
<td>TcII</td>
<td>Bolivia</td>
<td>Bolivia</td>
<td>Cochabamba</td>
</tr>
<tr>
<td>6406</td>
<td>M</td>
<td>67</td>
<td>Formosa, Goias</td>
<td>Megacolon, Megaesophagus</td>
<td>TcII</td>
<td>Brazil</td>
<td>Brazil</td>
<td>Goias</td>
</tr>
<tr>
<td>6407</td>
<td>F</td>
<td>40</td>
<td>MaraRosa, Goias</td>
<td>Cardiopathy</td>
<td>TcII</td>
<td>Brazil</td>
<td>Brazil</td>
<td>Goias</td>
</tr>
<tr>
<td>6408</td>
<td>M</td>
<td>38</td>
<td>Correntina, Bahia</td>
<td>Megaesophagus</td>
<td>TcII</td>
<td>Brazil</td>
<td>Brazil</td>
<td>Bahia</td>
</tr>
<tr>
<td>6409</td>
<td>F</td>
<td>54</td>
<td>Sao Luiz MBelos, Goias</td>
<td>Asymptomatic</td>
<td>TcII</td>
<td>Brazil</td>
<td>Brazil</td>
<td>Goias</td>
</tr>
<tr>
<td>6410</td>
<td>F</td>
<td>59</td>
<td>Brazabrantes, Goias</td>
<td>Asymptomatic</td>
<td>TcII</td>
<td>Brazil</td>
<td>Brazil</td>
<td>Goias</td>
</tr>
<tr>
<td>6411</td>
<td>M</td>
<td>56</td>
<td>Jussara, Goias</td>
<td>Megaesophagus</td>
<td>TcII</td>
<td>Brazil</td>
<td>Brazil</td>
<td>Goias</td>
</tr>
<tr>
<td>6412</td>
<td>M</td>
<td>30</td>
<td>SantaMariaVitoria, Bahia</td>
<td>Cardiopathy</td>
<td>TcII</td>
<td>Brazil</td>
<td>Brazil</td>
<td>Bahia</td>
</tr>
<tr>
<td>6413</td>
<td>F</td>
<td>24</td>
<td>Cocos, Bahia</td>
<td>Asymptomatic</td>
<td>TcII</td>
<td>Brazil</td>
<td>Brazil</td>
<td>Bahia</td>
</tr>
<tr>
<td>6414</td>
<td>M</td>
<td>32</td>
<td>Cocos, Bahia</td>
<td>Cardiopathy</td>
<td>TcII</td>
<td>Brazil</td>
<td>Brazil</td>
<td>Bahia</td>
</tr>
<tr>
<td>6415</td>
<td>F</td>
<td>47</td>
<td>Correntina, Bahia</td>
<td>Asymptomatic</td>
<td>TcII</td>
<td>Brazil</td>
<td>Brazil</td>
<td>Bahia</td>
</tr>
<tr>
<td>6416</td>
<td>M</td>
<td>30</td>
<td>SantaMariaVitoria, Bahia</td>
<td>Cardiopathy</td>
<td>TcII</td>
<td>Brazil</td>
<td>Brazil</td>
<td>Bahia</td>
</tr>
<tr>
<td>6417</td>
<td>F</td>
<td>47</td>
<td>Correntina, Bahia</td>
<td>Cardiopathy</td>
<td>TcII</td>
<td>Brazil</td>
<td>Brazil</td>
<td>Bahia</td>
</tr>
<tr>
<td>6418</td>
<td>F</td>
<td>24</td>
<td>Quilombo, Comitandia, Goias</td>
<td>Asymptomatic</td>
<td>TcII</td>
<td>Brazil</td>
<td>Brazil</td>
<td>Goias</td>
</tr>
<tr>
<td>ID</td>
<td>Sex</td>
<td>Age</td>
<td>Location</td>
<td>Disease</td>
<td>TcII/TcI</td>
<td>TCII</td>
<td></td>
<td></td>
</tr>
<tr>
<td>-----</td>
<td>-----</td>
<td>-----</td>
<td>-------------------------------</td>
<td>----------------------------------</td>
<td>-----------</td>
<td>--------</td>
<td></td>
<td></td>
</tr>
<tr>
<td>6407</td>
<td>F</td>
<td>52</td>
<td>Brazil</td>
<td>Cardiopathy, severe cardiopathy</td>
<td>0.2489</td>
<td>1.2966</td>
<td></td>
<td></td>
</tr>
<tr>
<td>6416</td>
<td>M</td>
<td>45</td>
<td>Brazil</td>
<td>Severe cardiopathy</td>
<td>0.2209</td>
<td>1.4904</td>
<td></td>
<td></td>
</tr>
<tr>
<td>6445</td>
<td>M</td>
<td>40</td>
<td>Brazil</td>
<td>Cardiopathy, TcII</td>
<td>0.7339</td>
<td>1.7071</td>
<td></td>
<td></td>
</tr>
<tr>
<td>6452</td>
<td>M</td>
<td>46</td>
<td>Brazil</td>
<td>Severe cardiopathy</td>
<td>0.2253</td>
<td>1.3164</td>
<td></td>
<td></td>
</tr>
<tr>
<td>6536</td>
<td>M</td>
<td>30</td>
<td>Brazil</td>
<td>Megaesophagus</td>
<td>0.5157</td>
<td>2.5105</td>
<td></td>
<td></td>
</tr>
<tr>
<td>6548</td>
<td>F</td>
<td>57</td>
<td>Brazil</td>
<td>Asymptomatic</td>
<td>ND</td>
<td>ND</td>
<td></td>
<td></td>
</tr>
<tr>
<td>6561</td>
<td>F</td>
<td>46</td>
<td>Brazil</td>
<td>Asymptomatic</td>
<td>ND</td>
<td>ND</td>
<td></td>
<td></td>
</tr>
<tr>
<td>6563</td>
<td>F</td>
<td>58</td>
<td>Brazil</td>
<td>Cardiopathy, Megaesophagus</td>
<td>ND</td>
<td>ND</td>
<td></td>
<td></td>
</tr>
<tr>
<td>6569</td>
<td>M</td>
<td>65</td>
<td>Brazil</td>
<td>Megaesophagus</td>
<td>0.3661</td>
<td>2.6595</td>
<td></td>
<td></td>
</tr>
<tr>
<td>6571</td>
<td>F</td>
<td>69</td>
<td>Brazil</td>
<td>Cardiopathy, Megaesophagus</td>
<td>ND</td>
<td>ND</td>
<td></td>
<td></td>
</tr>
<tr>
<td>6577</td>
<td>F</td>
<td>57</td>
<td>Brazil</td>
<td>Cardiopathy</td>
<td>ND</td>
<td>ND</td>
<td></td>
<td></td>
</tr>
<tr>
<td>6581</td>
<td>M</td>
<td>55</td>
<td>Brazil</td>
<td>Cardiopathy</td>
<td>0.1979</td>
<td>0.7195</td>
<td></td>
<td></td>
</tr>
<tr>
<td>6582</td>
<td>M</td>
<td>51</td>
<td>Brazil</td>
<td>Cardiopathy</td>
<td>0.3245</td>
<td>2.4802</td>
<td></td>
<td></td>
</tr>
<tr>
<td>6588</td>
<td>M</td>
<td>51</td>
<td>Brazil</td>
<td>Cardiopathy</td>
<td>1.0797</td>
<td>1.9403</td>
<td></td>
<td></td>
</tr>
<tr>
<td>6590</td>
<td>M</td>
<td>49</td>
<td>Brazil</td>
<td>Severe cardiopathy</td>
<td>0.3245</td>
<td>2.4802</td>
<td></td>
<td></td>
</tr>
<tr>
<td>6597</td>
<td>F</td>
<td>56</td>
<td>Brazil</td>
<td>Cardiopathy</td>
<td>0.4033</td>
<td>1.9877</td>
<td></td>
<td></td>
</tr>
<tr>
<td>6603</td>
<td>F</td>
<td>36</td>
<td>Brazil</td>
<td>Cardiopathy</td>
<td>0.4033</td>
<td>1.9877</td>
<td></td>
<td></td>
</tr>
<tr>
<td>6687</td>
<td>F</td>
<td>49</td>
<td>Brazil</td>
<td>Cardiopathy</td>
<td>0.3905</td>
<td>1.7523</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Index</td>
<td>ID</td>
<td>Sex</td>
<td>Age</td>
<td>Diagnosis</td>
<td>Country</td>
<td>City</td>
<td></td>
<td></td>
</tr>
<tr>
<td>-------</td>
<td>------</td>
<td>-----</td>
<td>-----</td>
<td>--------------------</td>
<td>---------</td>
<td>------</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>1,9928</td>
<td></td>
<td></td>
<td>Congenital</td>
<td>Bolivia</td>
<td>Sucre</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>2,8809</td>
<td></td>
<td></td>
<td>Congenital</td>
<td>Bolivia</td>
<td>Sucre</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>1,6921</td>
<td></td>
<td></td>
<td>Congenital</td>
<td>Bolivia</td>
<td>Sucre</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>2,0828</td>
<td></td>
<td></td>
<td>Congenital</td>
<td>Bolivia</td>
<td>Sucre</td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>2,3075</td>
<td></td>
<td></td>
<td>Congenital</td>
<td>Bolivia</td>
<td>Sucre</td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>1,5698</td>
<td></td>
<td></td>
<td>Congenital</td>
<td>Bolivia</td>
<td>Sucre</td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>1,7856</td>
<td></td>
<td></td>
<td>Congenital</td>
<td>Bolivia</td>
<td>Sucre</td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>1,3492</td>
<td></td>
<td></td>
<td>Congenital</td>
<td>Bolivia</td>
<td>Sucre</td>
<td></td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>1,6215</td>
<td></td>
<td></td>
<td>Congenital</td>
<td>Bolivia</td>
<td>Sucre</td>
<td></td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>1,8906</td>
<td></td>
<td></td>
<td>Congenital</td>
<td>Bolivia</td>
<td>Sucre</td>
<td></td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>1,3074</td>
<td></td>
<td></td>
<td>Congenital</td>
<td>Bolivia</td>
<td>Sucre</td>
<td></td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>1,6495</td>
<td></td>
<td></td>
<td>Congenital</td>
<td>Bolivia</td>
<td>Sucre</td>
<td></td>
<td></td>
</tr>
<tr>
<td>13</td>
<td>1,5264</td>
<td></td>
<td></td>
<td>Congenital</td>
<td>Bolivia</td>
<td>Sucre</td>
<td></td>
<td></td>
</tr>
<tr>
<td>14</td>
<td>1,6647</td>
<td></td>
<td></td>
<td>Congenital</td>
<td>Bolivia</td>
<td>Sucre</td>
<td></td>
<td></td>
</tr>
<tr>
<td>15</td>
<td>1,4771</td>
<td></td>
<td></td>
<td>Congenital</td>
<td>Bolivia</td>
<td>Sucre</td>
<td></td>
<td></td>
</tr>
<tr>
<td>16</td>
<td>1,1428</td>
<td></td>
<td></td>
<td>Congenital</td>
<td>Bolivia</td>
<td>Sucre</td>
<td></td>
<td></td>
</tr>
<tr>
<td>17</td>
<td>1,3370</td>
<td></td>
<td></td>
<td>Congenital</td>
<td>Bolivia</td>
<td>Sucre</td>
<td></td>
<td></td>
</tr>
<tr>
<td>18</td>
<td>1,2816</td>
<td></td>
<td></td>
<td>Congenital</td>
<td>Bolivia</td>
<td>Sucre</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*Note: The table above represents clinical data of patients with various diagnoses, including congenital conditions. The columns include index, ID, sex, age, diagnosis, country, and city.*
<table>
<thead>
<tr>
<th>Sample ID</th>
<th>F</th>
<th>TcIII-VI</th>
<th>Diagnosis</th>
<th>Location</th>
<th>Age</th>
<th>Sex</th>
<th>Distance Apart</th>
<th>Comments</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.716449</td>
<td>1.983814</td>
<td>ND</td>
<td>Congenital</td>
<td>Vinto, Bolivia</td>
<td>18</td>
<td>F</td>
<td>&gt;1 year</td>
<td>Samples from the same patient &gt;12 months apart</td>
</tr>
<tr>
<td>1.637314</td>
<td>1.716449</td>
<td>ND</td>
<td>Congenital</td>
<td>Vinto, Bolivia</td>
<td>&gt;1</td>
<td>ND</td>
<td>&lt;6 months</td>
<td>Samples from the same patient taken &lt;6 months apart</td>
</tr>
<tr>
<td>1.637314</td>
<td>1.716449</td>
<td>ND</td>
<td>Congenital</td>
<td>Vinto, Bolivia</td>
<td>&gt;1</td>
<td>ND</td>
<td>&gt;1 year</td>
<td>Samples from the same patient &gt;12 months apart</td>
</tr>
</tbody>
</table>

*CIUF* refers to the institute or organization conducting the study.
Table 2 - Yang and Neilson estimates for positive selection within and among abundant 97% STs identified in this study.

<table>
<thead>
<tr>
<th>Population</th>
<th>ST</th>
<th>Infecting strain</th>
<th>Sequences</th>
<th>S-S segregations</th>
<th>N-S segregations</th>
<th>S-S substitutions</th>
<th>N-S substitutions</th>
<th>Set</th>
<th>Method</th>
<th>KA</th>
<th>KS</th>
<th>KA/KS</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Goias</td>
<td>TcII</td>
<td>357227 (271)</td>
<td>0.06</td>
<td>0.07</td>
<td>0.8354</td>
<td>0.0000</td>
<td>0.0000</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.0000</td>
</tr>
<tr>
<td></td>
<td>TcII</td>
<td>236805 (149)</td>
<td>7.45</td>
<td>2.82</td>
<td>2.6436</td>
<td>0.0000</td>
<td>0.0000</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.0000</td>
</tr>
<tr>
<td></td>
<td>TcII</td>
<td>96274 (82)</td>
<td>7.04</td>
<td>1.11</td>
<td>6.3415</td>
<td>0.0000</td>
<td>0.0000</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.0000</td>
</tr>
<tr>
<td></td>
<td>TcII</td>
<td>9981 (19)</td>
<td>0.02</td>
<td>0.05</td>
<td>0.4151</td>
<td>0.0000</td>
<td>0.0000</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.0000</td>
</tr>
<tr>
<td>Bolivia</td>
<td>TcV/TcI</td>
<td>59431 (86)</td>
<td>0.06</td>
<td>0.08</td>
<td>0.7515</td>
<td>0.0002</td>
<td>0.0000</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.0000</td>
</tr>
<tr>
<td></td>
<td>TcV</td>
<td>38455 (36)</td>
<td>0.02</td>
<td>0.03</td>
<td>0.7876</td>
<td>0.1600</td>
<td>0.0000</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.1600</td>
</tr>
<tr>
<td></td>
<td>TcV</td>
<td>12676 (24)</td>
<td>0.03</td>
<td>0.03</td>
<td>0.7868</td>
<td>0.2290</td>
<td>0.0000</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.2290</td>
</tr>
<tr>
<td></td>
<td>TcI</td>
<td>3448 (13)</td>
<td>5.57</td>
<td>1.98</td>
<td>2.8059</td>
<td>0.0000</td>
<td>0.0000</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.0000</td>
</tr>
<tr>
<td></td>
<td>TcI</td>
<td>242 (3)</td>
<td>4.48</td>
<td>3.61</td>
<td>1.2422</td>
<td>0.3484</td>
<td>0.0000</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.3484</td>
</tr>
</tbody>
</table>

Numbers in brackets represent the number of 99% STs defined within each cluster from which estimates were generated.

P values are given for Fisher's exact tests for deviation from the neutral expectation of KA/KS = 0.
<table>
<thead>
<tr>
<th>Code</th>
<th>Description</th>
<th>Read/Depth</th>
<th>Number</th>
<th>Code</th>
<th>Description</th>
<th>Read/Depth</th>
<th>Number</th>
</tr>
</thead>
<tbody>
<tr>
<td>41</td>
<td>Bolivian Infant 1</td>
<td>1874</td>
<td>Goias</td>
<td>45</td>
<td>Bolivian Infant 1</td>
<td>1874</td>
<td>Goias</td>
</tr>
<tr>
<td>53</td>
<td>Bolivian Mother 1</td>
<td>1729</td>
<td></td>
<td>63</td>
<td>Bolivian Mother 1</td>
<td>1729</td>
<td></td>
</tr>
<tr>
<td>136</td>
<td>Bolivian Infant 2</td>
<td>8443</td>
<td></td>
<td>24</td>
<td>Bolivian Mother 2</td>
<td>8494</td>
<td></td>
</tr>
<tr>
<td>128</td>
<td>Bolivian Mother 2</td>
<td>8494</td>
<td></td>
<td>31</td>
<td>Bolivian Infant 2</td>
<td>8443</td>
<td></td>
</tr>
<tr>
<td>46</td>
<td>Bolivian Infant 3</td>
<td>2242</td>
<td></td>
<td>25</td>
<td>Bolivian Mother 3</td>
<td>1871</td>
<td></td>
</tr>
<tr>
<td>35</td>
<td>Bolivian Mother 3</td>
<td>1871</td>
<td></td>
<td>40</td>
<td>Bolivian Infant 3</td>
<td>2242</td>
<td></td>
</tr>
<tr>
<td>60</td>
<td>Bolivian Infant 4</td>
<td>1368</td>
<td></td>
<td>42</td>
<td>Bolivian Mother 4</td>
<td>1367</td>
<td></td>
</tr>
<tr>
<td>32</td>
<td>Bolivian Mother 4</td>
<td>1367</td>
<td></td>
<td>65</td>
<td>Bolivian Infant 4</td>
<td>1368</td>
<td></td>
</tr>
<tr>
<td>53</td>
<td>Bolivian Infant 5</td>
<td>1411</td>
<td></td>
<td>26</td>
<td>Bolivian Mother 5</td>
<td>1387</td>
<td></td>
</tr>
<tr>
<td>48</td>
<td>Bolivian Mother 5</td>
<td>1387</td>
<td></td>
<td>105</td>
<td>Bolivian Infant 5</td>
<td>1411</td>
<td></td>
</tr>
<tr>
<td>284</td>
<td>Bolivian Infant 6</td>
<td>5184</td>
<td></td>
<td>104</td>
<td>Bolivian Mother 6</td>
<td>5139</td>
<td></td>
</tr>
<tr>
<td>234</td>
<td>Bolivian Mother 6</td>
<td>5139</td>
<td></td>
<td>105</td>
<td>Bolivian Infant 6</td>
<td>5184</td>
<td></td>
</tr>
<tr>
<td>59</td>
<td>Bolivian Infant 7</td>
<td>1531</td>
<td></td>
<td>75</td>
<td>Bolivian Mother 7</td>
<td>1532</td>
<td></td>
</tr>
<tr>
<td>154</td>
<td>Bolivian Mother 7</td>
<td>1532</td>
<td></td>
<td>35</td>
<td>Bolivian Infant 7</td>
<td>1531</td>
<td></td>
</tr>
<tr>
<td>69</td>
<td>Bolivian Infant 8</td>
<td>1241</td>
<td></td>
<td>35</td>
<td>Bolivian Mother 8</td>
<td>1242</td>
<td></td>
</tr>
<tr>
<td>63</td>
<td>Bolivian Mother 8</td>
<td>1242</td>
<td></td>
<td>93</td>
<td>Bolivian Infant 8</td>
<td>1241</td>
<td></td>
</tr>
<tr>
<td>127</td>
<td>Bolivian Infant 9</td>
<td>2900</td>
<td></td>
<td>84</td>
<td>Bolivian Mother 9</td>
<td>2832</td>
<td></td>
</tr>
<tr>
<td>72</td>
<td>Bolivian Mother 9</td>
<td>2832</td>
<td></td>
<td>91</td>
<td>Bolivian Infant 9</td>
<td>2900</td>
<td></td>
</tr>
<tr>
<td>107</td>
<td>Bolivian Infant 10</td>
<td>2076</td>
<td></td>
<td>98</td>
<td>Bolivian Mother 10</td>
<td>2144</td>
<td></td>
</tr>
<tr>
<td>128</td>
<td>Bolivian Mother 10</td>
<td>2144</td>
<td></td>
<td>108</td>
<td>Bolivian Infant 10</td>
<td>2076</td>
<td></td>
</tr>
<tr>
<td>6720</td>
<td>Goias Infant 1</td>
<td>7661</td>
<td></td>
<td>75</td>
<td>Goias Mother 1</td>
<td>8193</td>
<td></td>
</tr>
<tr>
<td>311</td>
<td>Goias Mother 1</td>
<td>8193</td>
<td></td>
<td>128</td>
<td></td>
<td>6718</td>
<td></td>
</tr>
</tbody>
</table>
Figure 4
3.4 Discussion

In lieu of comparative genomics of representative *T. cruzi* field isolates, not yet a reality, as is the case with other more experimentally-tractable trypanosomatids (Downing et al., 2011; Goodhead et al., 2013; Rogers et al., 2014), the four papers included in this chapter describe recent advances in *T. cruzi* genotyping technology.

Multiple, single-copy, chromosomally-independent, nuclear housekeeping genes were evaluated initially for reliable amplification, their ability to assign isolates to DTU-level, to facilitate higher resolution intra-lineage analyses (Yeo et al., 2011) and finally their inclusion alongside additional targets in a standardized *T. cruzi* MLST scheme (Diosque et al., 2014). Overall nuclear MLST (nMLST) was a robust, reproducible, and in many cases, highly discriminatory method to characterize *T. cruzi* isolates to DTU-level. Modified data were amenable to statistical algorithms available for bacterial MLST (e.g. eBURST (Feil et al., 2004)) and now new software has been developed specifically to handle diploid eukaryotic sequence data (Tomasini et al., 2013).

Potential limitations of this technique, highlighted by these studies, include unpredictable LOH observed among different strains and loci, which resulted in spurious phylogenetic incongruence and isolate misclassification (especially when loci were examined individually). However, this phenomenon is interesting to examine in the context of gene conversion and homologous recombination, which have both been implicated in inflating homozygosity estimates in *T. cruzi* previously (Llewellyn et al., 2009a; 2011). Additionally, candidate genes were assessed against a limited number of reference isolates (39 and 25, respectively), which encompass current known *T. cruzi* genetic diversity, but were under-representative of some particular disease foci, e.g. Colombia and Central America.

For the immediate future, these promising results argue for nMLST to become the new gold standard for DTU assignment of isolated parasite strains. A minimum panel of four targets can allocate samples to DTU-level, while up to seven loci afforded potential intra-lineage resolution. For future applications, it is likely that the combination and number of markers used will depend on the particular strain origin and DTU(s) under investigation. For example, recent nMLST studies of TcI isolates from geographically-restricted areas of the Argentinian Chaco and North-East Colombia relied on eight and thirteen loci, respectively, to obtain sufficient intra-lineage resolution (Tomasini et al., 2014a; Ramirez et al., 2013d).

An alternate approach for intra-DTU studies in *T. cruzi* is analysis of more variable markers such as maxicircle genes or microsatellites which both have mutation rates several orders of magnitude higher than protein coding genes (Ellegren, 2000; Ballard and Whitlock, 2004; Lewis et al., 2011). Firstly, to address undefined intra-strain homology associated with multi-copy mitochondrial markers, Illumina sequencing reads, generated as part of the Sylvio X10/1 genome project (Franzen et al., 2011), were reassembled and mapped against the maxicircle genome, to demonstrate that heteroplasmy within the coding region was present at <0.5% of sites. Based on these observations, an MLST scheme comprising ten maxicircle gene fragments was developed and evaluated for reproducible amplification across the six *T. cruzi* DTUs and subsequently for its levels of intra-TcI resolution, compared to conventional
single (GPI and SL-IR) and multiple (MLMT) nuclear genotyping markers (Messenger et al., 2012).

One advantage of this combined nuclear-mitochondrial methodology over nMLST alone, is its ability to detect cryptic hybridization and identify potential underlying mechanisms (in this study, seemingly asymmetric mitochondrial introgression with minimal reciprocal nuclear recombination) (Messenger et al., 2012). However, these observations also caution the interpretation of mtMLST without any comparative nuclear data, which could lend itself to incorrect strain assignment. Similarly, phylogenetic incongruence between nMLST topologies may also be indicative of historical recombination but, at least in T. cruzi, appears to occur less frequently (Yeo et al., 2011; Diosque et al., 2014; Tomasini et al., 2014). The genotyping approach described by Messenger et al., 2012, is currently recommended for intra-lineage population genetics studies (Ramírez et al., 2012; Zumaya-Estrada et al., 2012; Lima et al., 2014) and can also be expanded to include nMLST (see chapter 5), enabling analysis of three different types of molecular marker at several overlapping levels of resolution.

The application of MLST (both nuclear and mitochondria) in endemic areas, is restricted primarily by reagent cost, technical expertise and access to a sequencer. The same limitations apply to other genotyping methods, such as microsatellites, which are notoriously difficult to reproduce precisely between laboratories. Without concomitant improvements in infrastructure, it is evident that in some of the more underdeveloped regions, routine genotyping of T. cruzi will still need to rely on lower technological approaches e.g. PCR-RFLPs (Rozas et al., 2007; Lewis et al., 2009a; Cosentino and Agüero, 2012).

Finally, Illumina amplicon deep sequencing markers were developed to examine parasite multiclonality within individual chronic and congenital chagasic patients; putative evidence of diversifying selection affecting antigenic genes was observed, suggesting a link between genetic diversity in this gene family and survival in the mammalian host (Llewellyn et al., accepted). The failure of this study to identify any correlation between parasite genetic diversity and patient sex, age or clinical symptoms, may simply reflect limitations of this particular sample cohort. The ‘ideal’ population to examine such associations is an endemic community where active vector transmission and high force of infection persists, such as those described in the Bolivian Gran Chaco, where Chagas seroprevalence is directly proportional to age, reaching up to 97% in adults older than 30 years (Samuels et al., 2013). While it is generally accepted that geographical variation in chronic Chagas disease pathology is partially attributable to differences in parasite genetic diversity, no clear relationship between T. cruzi genotype and clinical disease status has yet been established (Miles et al., 2009).

With targeted cross-sectional sampling of such a population, deep sequencing has the potential to investigate phenomena such as super-infection, which may be more relevant in the context of Chagas disease pathogenesis. Super-infection due to ongoing vector exposure has been proposed to increase parasite load, sustain antigen exposure and consequent inflammatory immune response at a higher levels, thereby driving progression of cardiac
morbidity and risk of congenital transmission (Bustamante et al., 2002; 2003; 2004; 2007; Andrade et al., 2006; Torrico et al., 2006; Marin-Neto et al., 2007; Dutra and Gollob, 2008). Future studies could adopt a multidisciplinary approach, incorporating parallel immunological assays and quantification of parasitaemia by qPCR to investigate the impact of T. cruzi multiclonality on chronic immune dysregulation. Additionally, improvements can also be made with regards to sequencing technology, particularly elimination of any PCR steps which likely biased ST recovery, and exploration of deep sequencing directly from clinical samples, e.g. via whole genome amplification (WGA), to circumvent potential loss of clonal diversity during parasite enrichment stages (Morrison et al., 2007).

All of the genotyping methods described herein supported the six current T. cruzi DTUs. By comparison the level of nuclear sequence divergence between major lineages was equivalent to inter-species diversity among New World Leishmania species (Yeo et al., 2011; Boité et al., 2012), raising the question, is current nomenclature sufficient to describe known T. cruzi genetic diversity? There appears to be mounting evidence for the inclusion of a seventh DTU: TcBat (Marcili et al., 2009; Pinto et al., 2012; Ramírez et al., 2014). No TcBat isolates were examined as part of these studies. It would be interesting to apply these markers to the genetic characterization of this novel lineage, to confirm its affiliations with TcI, and investigate its proposed role as the antecedent of the other T. cruzi DTUs (Guhl et al., 2014).

Recently, there has been significant interest in subdividing TcI by transmission cycle (Tcla-e) (Herrera et al., 2007b; 2009; Falla et al., 2009; Cura et al., 2010; Guhl and Ramírez, 2011) on the basis of variability in the SL-IR. However, when multiple MLMT and mtMLST markers have been applied to equivalent strains, these subdivisions, particularly among peridomestic and sylvatic cycles, collapse. Instead TcI isolates group either alongside low diversity isolates from humans and domestic vectors (TcIDOM) or with sylvatic samples that display continental-wide spatial structuring (Llewellyn et al., 2009a; Zumaya-Estrada et al., 2012). The latter designation of a subset of human-associated genotypes is arguably more biologically relevant, but may still be characterizing diversity for its own sake.

It is probable that as more T. cruzi whole genome sequences become available additional distinct genotypes relating to pathogenesis, including structural and copy number variants (Urban et al., 2011; Minning et al., 2011; Sterkers et al., 2011; Pavia et al., 2012), may be uncovered and call for nomenclature reclassification. Until that time, presented herein are some of the highest resolution genotyping techniques developed in T. cruzi to date which support and will potentially expand our current understanding of parasite genetic diversity.
4. Investigating the association between *T. cruzi* genetic diversity and ecophylogeography of Chagas disease

4.1 Background
An improved understanding of the interactions between natural parasite populations and their environment is crucial to establish the epidemiological risk associated with emergent pathogenic genotypes. *T. cruzi* is an ancient, pervasive multi-host zoonotic disease which was likely introduced to South America via North American bats approximately 7-10 MYA (Stadelmann et al., 2007). The diversification of *T. cruzi* from its MRCA into its current DTUs occurred within the last 3-4 MYA (Flores-López and Machado, 2011; Lewis et al., 2011).

Paleoparasitology data indicate that soon after colonizing South America ~15,000 years ago (Goebel et al., 2008), humans became infected with *T. cruzi*; the earliest recorded human infection is from a 9000 year old Chinchorro mummy, which inhabited the coastal region of the Atacama Desert (Aufderheide et al., 2004). Similar incidences of *T. cruzi* infection among mummies from subsequent cultures that succeeded the Chinchorros (including the Alto Ramirez, Cabuza, Maitas, Chiribaya, San Miguel, Inca and Colonial people), at an average prevalence of 40.6%, suggests that by pre-Colombian times, Chagas disease was already widespread in human societies (Aufderheide et al., 2004). Subsequent studies in the same area have identified different *T. cruzi* genotypes predominating throughout this time period; TcI and TcII among the Chinchorros were replaced by TcV and TcVI among Cabuza and Chiribaya mummies (Guhl et al., 2014). TcI and TcII infections have also been detected in more recent human remains from other archeological sites in Minas Gerais, Brazil (Fernandes et al., 2008; Lima et al., 2008) and the Chihuahuan Desert, Texas (Dittmar et al., 2003; Araújo et al., 2009), respectively.

Based on these observations it has been hypothesized that Chagas disease originated in the Andean region amongst the nomadic Chinchorro people (Ferreira et al., 2011) whose descendants were the first to domesticate wild guinea pigs (likely *Cavia tschudi*) for consumption and religious rituals (Gade, 1967; Wing, 1986), initiated maize storage as part of early arable farming (Rothhammer et al., 1985) and undertook ritual pilgrimages (Cortez et al., 2010), thereby providing ample habitats, bloodmeal sources and dispersal mechanisms for *T. infestans*, also believed to have emerged and become domiciliated in the Bolivian highlands (Dujardin, 1998; Panzera et al., 2004; Bargues et al., 2006; Cortez et al., 2010).

In parallel, *T. cruzi* had already dispersed throughout the sylvatic environment, adapting to infect a wide range of mammalian reservoirs including members of the orders *Marsupialia, Xenarthra, Rodentia, Primates, Carnivora, Chiroptera* and *Artiodactyla* (Noireau et al., 2009a; Jansen and Roque, 2010). As yet, relatively few in depth host association studies between *T. cruzi* and its triatomine vectors and vertebrate hosts have been conducted. TcI is frequently isolated from the triatomine tribe *Rhodniini*, and arboreal marsupials, particularly *D. marsupialis* and TcIII from *Dasypus* species (Gaunt and Miles, 2000; Yeo et al., 2005;
Llewellyn et al., 2009a; 2009b). However, in both cases these relationships are not absolute and genotypes appear to cluster according to geography rather than host species (Acosta et al., 2001; Llewellyn et al., 2009a; 2009b; Marcili et al., 2009b).

Interestingly, animal models suggest that not all mammalian host species are equally susceptible to infection by different *T. cruzi* DTUs. Attempts to artificially infect two species of opossum (*Monodelphis domestica* and *Didelphis virginiana*) demonstrated that both were resistant to TcIV but fully susceptible to TcI (Roellig et al., 2009; 2010). Similar observations have been reported with *Rhodnius* vectors which tend to be less permissive to experimental TcII infection (Mello et al., 1996; Araújo et al., 2014). By comparison, recent field evidence from Brazil have proposed that some *Carnivora* can act as indiscriminant bioaccumulators of *T. cruzi* genetic diversity with the potential to connect independent transmission cycles (Rocha et al., 2013). The key determinants of contemporary sylvatic host associations and parasite diversification are largely unknown.

Also of particular interest in the context of sylvatic *T. cruzi* transmission, is the impact of anthropogenic activity on the ecoepidemiology of Chagas disease and the evolution of *T. cruzi*. Habitat destruction has the capacity to drive triatomine species to invade the domestic environment; successful control of one species in houses can expose a niche for others to invade, especially in areas such as Venezuela, where domestic and sylvatic transmission cycles potentially overlap (Fitzpatrick et al., 2008). Likewise, in areas unaffected by deforestation, including some of the indigenous inhabitations of the Amazon, domestic *T. cruzi* transmission has thus far failed to establish (Miles et al., 1978; Walsh et al., 1993).

Human encroachment into sylvatic areas has undoubtedly influenced parasite genetic diversity and has been hypothesized to explain the emergence of TcV and TcVI and their seemingly epidemic propagation across the Southern Cone (Lewis et al., 2011). Similarly, multiple molecular markers suggest that TcI human infections in Venezuela and Colombia are associated with genetically homogenous strains (TcI<sub>DOM</sub>/TcI<sub>a</sub>/VEN<sub>DOM</sub>) which are distinct and absent from local sylvatic populations (Herrera et al., 2007b; 2009; O’Connor et al., 2007; Falla et al., 2009; Llewellyn et al., 2009a; Ocaña-Mayorga et al., 2010; Cura et al., 2010; Ramírez et al., 2012). The pathological implications of these low diversity genotypes and their evolutionary origins remain unresolved.
4.2 Objectives
The aim of this chapter was to exploit the phylogenetic markers developed in chapter 3 to investigate the interaction between parasite genetic diversity and ecophylogeography of Chagas disease.

Specific objectives were to:

a. Identify ecological determinants of sylvatic Tcl genetic diversification in arboreal and terrestrial transmission cycles in Bolivia.

b. Examine putative hybrid contact zones to detect incidences of nuclear-mitochondrial phylogenetic incongruence, which are indicative of natural genetic exchange.

c. Explore the interaction between intra-Tcl genetic heterogeneity and ecological biodiversity among different biomes in Brazil.

d. Investigate the evolutionary origin of TclDOM, a subset of homogeneous strains associated with human infections in northern South and Central America.
Results

4.3.1 Ecological host fitting of sylvatic T. cruzi

Two hypotheses exist to explain contemporary T. cruzi host associations. Some have proposed that T. cruzi co-evolved in close synergy with discrete vertebrate hosts and insect vectors (Miles et al., 1981b; Gaunt and Miles, 2000; Yeo et al., 2005), while others favour ecological host fitting (Hamilton et al., 2007; Llewellyn et al., 2009a; 2009b), i.e. co-option of existing genetic traits to colonize an unfamiliar resource, environmental niche or facilitate host switching. To identify key determinants of sylvatic T. cruzi genetic diversification and explore potential hybridization and spatial genetic structure of natural parasite populations, high resolution nuclear (Llewellyn et al., 2009a) and mitochondrial (Messenger et al., 2012) genotyping was used to characterize 199 contemporaneous TcI clones isolated from sylvatic transmission cycles in Bolivia.

This study is reported in full below in Messenger et al. accepted.

In summary:

- Based on multiple clustering algorithms (a non-parametric discriminant analysis of principal components (DAPC) and a NJ-DAS tree), TcI clones were grouped into three distinct sylvatic transmission cycles, corresponding to one terrestrial highland population (Cochabamba, Tupiza and Toro Toro), composed of genetically homogenous strains ($A_r = 1.92-2.22$; $PA/L = 0.19-0.42$; $D_{AS} = 0.151$; $Hd=0.54$) and two adjacent, highly diverse, arboreal lowland groups (East and North Beni) ($A_r = 3.40$ and $3.93$; $PA/L = 1.12$ and $0.60$ $D_{AS} = 0.425$ and $0.311$; $Hd=0.84$ and $0.81$, respectively).

- Estimates of subdivision ($F_{ST}$) demonstrated limited gene flow between neighbouring terrestrial and arboreal areas (Cochabamba-Beni distance ~220 km; $F_{ST} = 0.42$ and $0.35$) and low levels of subdivision among similar but geographically-disparate terrestrial ecotopes (Cochabamba-Tupiza distance >465 km; $F_{ST} = 0.016-0.084$) and adjacent arboreal study sites (North – East Beni distance = ~155km; $F_{ST} = 0.087$). A hierarchical AMOVA, indicated 23% of genetic diversity was attributed to differences between highland and lowlands, while only 4.5% and 7% were present at the population and the individuals between populations levels, respectively. Together these observations strongly support ecological host fitting as the predominant mechanism structuring parasite populations.

- Potential differences in mating strategies between highland and lowland populations was evidenced by dissimilar heterozygosity estimates (excess: $F_{IS} = -0.241$-$0.026$, 5-13% polymorphic loci with significant deficit in heterozygosity, and deficit: $F_{IS} = 0.176$ and $0.203$, 63.2% and 52.3% polymorphic loci with significant deficit in heterozygosity, respectively) and mitochondrial introgression among lowland strains. Significant nuclear-mitochondrial phylogenetic incongruence (SH test: ML tree $L = -4845.86$, Bayesian tree $L = -4849.55$ vs. $D_{AS}$ tree $L = -5006.48$, $p = 0.001$) was indicative of multiple independent genetic exchange events occurring in a potential hybridization zone in East Beni. A second contact area was putatively identified in Chapare, North Cochabamba, where a
single terrestrial isolate was more closely related to lowland populations by both nuclear and mitochondrial loci. Additionally, human isolates from Cochabamba, while genetically distinct from sylvatic strains circulating in the same area, were clustered together in the mitochondrial topology.

- Significant nuclear isolation by distance (IBD) was detected among lowland isolates ($R_{XY} = 0.209, p<0.001$, slope $= 0.0003 \pm 0.0000179$), while limited evidence for spatial structuring was apparent among highland strains from the same spatial scale ($R_{XY} = 0.109, p = 0.085$, slope $= 0.0002 \pm 0.00000873$). Concordant with $F_{ST}$ estimates between populations, these results support accelerated parasite dissemination between densely populated highland areas, compared to uninhabited lowland foci, which might be explained by passive, long-range anthroponotic dispersal, supported by evidence of mitochondrial gene flow between domestic and sylvatic populations in Cochabamba.
Please be aware that one cover sheet must be completed for each 'Research Paper' included in a thesis.

1. For a 'research paper' already published

1.1. Where was the work published? ..............................................................................................................................

1.2. When was the work published? ..............................................................................................................................

1.2.1. If the work was published prior to registration for your research degree, give a brief rationale for its inclusion
..............................................................................................................................................................................................................................
..............................................................................................................................................................................................................................
..............................................................................................................................................................................................................................

1.3. Was the work subject to academic peer review? .............................................................................................................

1.4. Have you retained the copyright for the work? Yes / No
   If yes, please attach evidence of retention.
   If no, or if the work is being included in its published format, please attach evidence of permission from copyright holder (publisher or other author) to include work

2. For a 'research paper' prepared for publication but not yet published

2.1. Where is the work intended to be published? Molecular Ecology
..............................................................................................................................................................................................................................

2.2. Please list the paper's authors in the intended authorship order

2.3. Stage of publication – Not yet submitted / Submitted / Undergoing revision from peer reviewers’ comments (In press)

3. For multi-authored work, give full details of your role in the research included in the paper and in the preparation of the paper. (Attach a further sheet if necessary)
   The candidate designed the study in close collaboration with Dr. Martin Llewellyn, derived the majority of biological clones analyzed, performed all of the genotyping experiments, analyzed the data and drafted the manuscript.

NAME IN FULL (Block Capitals) LOUISA ALEXANDRA MESSENGER

STUDENT ID NO: 223021

CANDIDATE’S SIGNATURE.................................................... Date 10/2/15

SUPERVISOR/SENIOR AUTHOR’S SIGNATURE (3 above) .............................................................................................................
Title: Ecological host fitting of Trypanosoma cruzi TcI in Bolivia: mosaic population structure, hybridization and a role for humans in Andean parasite dispersal.

Running Title: Ecological host fitting of T. cruzi TcI in Bolivia

Louisa A Messenger¹, Lineth Garcia², Mathieu Vanhove³, Carlos Huaranca², Marinely Bustamante², Marycruz Torrico², Faustino Torrico², Michael A Miles¹, Martin S Llewellyn¹

¹Department of Pathogen Molecular Biology, Faculty of Infectious and Tropical Diseases, London School of Hygiene and Tropical Medicine, London, UK
²Institute of Biomedical Research, Universidad Mayor de San Simón, Cochabamba, Bolivia
³Department of Infectious Disease Epidemiology, Imperial College London, London, UK

Abstract

An improved understanding of how a parasite species exploits its genetic repertoire to colonize novel hosts and environmental niches is crucial to establish the epidemiological risk associated with emergent pathogenic genotypes. Trypanosoma cruzi, a genetically heterogeneous, multi-host zoonosis, provides an ideal system to examine the sylvatic diversification of parasitic protozoa. In Bolivia T. cruzi I, the oldest and most widespread genetic lineage, is pervasive across a range of ecological clines. High resolution nuclear (26 loci) and mitochondrial (10 loci) genotyping of 199 contemporaneous sylvatic TcI clones was undertaken to provide insights into the biogeographical basis of T. cruzi evolution. Three distinct sylvatic transmission cycles were identified, corresponding to one terrestrial highland population, composed of genetically homogenous strains (Ar = 2.95; PA/L = 0.61; DAS = 0.151) and two adjacent, highly diverse, arboreal lowland groups (Ar = 3.40 and 3.93; PA/L = 1.12 and 0.60, DAS = 0.425 and 0.311, respectively). Limited gene flow between neighbouring terrestrial and arboreal areas (Distance ~220 km; FST = 0.42 and 0.35) and low levels of subdivision among similar but geographically-disparate terrestrial ecotopes (Distance >465 km; FST = 0.016-0.084) strongly support ecological host fitting as the predominant mechanism of parasite diversification. Dissimilar heterozygosity estimates (excess in highlands, deficit in lowlands) and mitochondrial introgression among lowland strains, may indicate fundamental differences in mating strategies between populations. Finally, accelerated parasite dissemination between densely populated, highland areas, compared to uninhabited lowland foci, likely reflect passive, long-range anthroponotic dispersal. The impact of humans on the risk of epizootic Chagas disease transmission in Bolivia is discussed.
Keywords

Trypanosoma cruzi, Population genetics, Microsatellites, Mitochondria, Sylvatic transmission, Ecological fitting

Introduction

Host-parasite relationships are assumed to be ecologically-specialised, tightly co-evolved systems driven by either mutual modification (‘synchronous co-speciation’) or exaptation into novel environmental niches, often accompanied by host switching (‘ecological fitting’) (Janzen 1985; Brooks et al. 2006). Ecological fitting occurs when an organism co-opts their existing suite of genetic traits to exploit an unfamiliar resource or colonize and persist in a new or modified environment (Agosta & Klemens 2008). Distinguishing between host-parasite relationships that result from ecological fitting or long-term co-evolution remains challenging. However, understanding how a species exploits their existing genetic repertoire to form novel host associations, is of primary interest to the study of emerging infectious diseases, with considerable implications for the design of disease control programmes (Brooks & Ferrao 2005; Agosta et al. 2010). In this regard, Trypanosoma cruzi (Kinetoplastida: Trypanosomatidae), the aetiological agent of Chagas disease, a pervasive zoonosis that is eclectic with respect to ectotope and host, provides a model system to examine the genetic diversification of parasitic protozoa.

Chagas disease is the most important vector-borne infection in Latin America, affecting an estimated 8-10 million individuals, with a further 90 million at risk (Hotez et al. 2008). Following acute disease, which is often undiagnosed, the majority of patients are clinically asymptomatic for life. Without treatment, approximately 20-30% will develop irreversible, potentially fatal cardiomyopathy, or more rarely, dilatation of the gastrointestinal tract (megaesophagus or megacolon) (Rassi Jr et al. 2010). The geographical distribution of T. cruzi extends from the southern United States to Argentinean Patagonia, where it is transmitted by more than 100 species of hematophagous triatomine bugs (Hemiptera: Reduviidae: Triatominae) (Lent & Wygodzinsky 1979; Galvão et al. 2003). Human disease is primarily confined to areas of Central and South America where individuals are exposed to infected faeces of domiciliated or invasive triatomines through contact with intact mucosae or abraded skin (Coura & Dias 2009). In addition, enzootic T. cruzi infection is naturally sustained by an extensive range of domestic, synanthropic and sylvatic mammalian hosts (Noireau et al. 2009).

T. cruzi is an ancient parasite, estimated to have diverged from its most recent common ancestor 3-4 million years ago (Lewis et al. 2011), and as such, is characterized by considerable genetic diversity (Stevens et al. 1999). Current international consensus recognises a minimum of six stable genetic lineages or discrete typing units (DTUs) (TcI-TcVI) (Zingales et al. 2009), which have...
distributions loosely defined by geography, ecology and transmission cycle (Miles et al. 2009). The level of nuclear sequence divergence between major *T. cruzi* DTUs is equivalent to inter-species diversity among New World *Leishmania* species (Boité et al. 2012; Yeo et al. 2011). TcI is the most widely distributed DTU; it is the principal cause of human chagasic cardiomyopathy in Colombia and Venezuela (Ramírez et al. 2010; Carrasco et al. 2012) and is ubiquitous among sylvatic transmission cycles across its endemic range (Llewellyn et al. 2009a). Multiple molecular markers consistently identify high levels of genetic diversity within sylvatic TcI populations, (Herrera et al. 2007; Herrera et al. 2009; O’Connor et al. 2007; Falla et al. 2009; Llewellyn et al. 2009a; Ocaña-Mayorga et al. 2010; Lima et al. 2014), and divergent, but genetically homogeneous, strains isolated from human infections (Llewellyn et al. 2009a; Cura et al. 2010; Ramírez et al. 2012; Zumaya-Estrada et al. 2012). However, the genetic determinants that drive natural *T. cruzi* diversification are largely unknown. Some have proposed that *T. cruzi* lineages co-evolved in close concert with discrete vertebrate hosts and insect vectors (Miles et al. 1981a; Gaunt & Miles 2000; Yeo et al. 2005), while others favour ecological fitting as a more parsimonious explanation for contemporary host associations (Hamilton et al. 2007; Agosta & Klemens 2008; Llewellyn et al. 2009a). Evidence to support the latter is increasing; TcI has now been isolated from the Orders Didelphimorphia, Rodentia., Carnivora and Primates, spanning multiple ecological niches (Lima et al. 2014; Rocha et al. 2013; Llewellyn et al. 2009a; Herrera et al. 2008a; Herrera et al. 2008b; Herrera et al. 2005; Lisboa et al. 2004), and genetic diversity of terrestrial TcIII appears similarly independent of host species (Llewellyn et al. 2009b; Marcilli et al. 2009).

Bolivia offers a range of diverse sylvatic ecotopes where *T. cruzi* transmission persists unabated. Colonies of *Triatoma infestans*, infected with TcI (Brenière et al. 2012), have been reported in highland Andean valleys (Buitrago et al. 2010; Cortez et al. 2006; Cortez et al. 2007) and to the South in the arid, lowland Chaco region (Waleckx et al. 2012; Ceballos et al. 2011), where their potential for domestic re-invasion threatens the success of the National Control Programme (Noireau et al. 2005; Noireau 2009). Sylvatic transmission also extends northwards to sparsely populated Amazonian Beni, where disease ecology is poorly described (Mattias et al. 2003; Justi et al. 2010). Human Chagas disease remains a prominent public health problem in Bolivia, affecting roughly 6.75% of the population (Jannin & Salvatella 2006). It is endemic across two-thirds of the country and concentrated disproportionally among lower socioeconomic rural populations with seroprevalence reaching 72.7-97.1% among adults of some communities (Medrano-Mercado et al. 2008; Samuels et al. 2013). Continuing domestic transmission, principally of TcI and TcV (Flores-Chavez et al. 2006; Barnabé et al. 2011; Bosseno et al. 1996), can be attributed to a decrease in intensity of residual insecticide spraying (Samuels et al. 2013; Espinoza et al. 2014), the emergence of insecticide resistance (Lardeux et al. 2010; Germano et al. 2010) and decentralized vector control initiatives in areas of recurrent political, social and economic instability (Gürtler 2009).
To date few studies have adopted rigorous sampling strategies and genetic markers with sufficient resolution to elucidate fully the biogeographical basis of *T. cruzi* evolution. Ideally, parasite samples should be minimally subdivided biologically, spatially and temporally, with multiple clones examined from each host (Prugnolle & De Meeus 2010). In practice, low circulating parasitaemia often prohibits parasite isolation and thus many studies are heavily reliant on historical collections of reference isolates. *T. cruzi* genetic analysis is further complicated by the presence of mixed DTU infections (Burgos et al. 2008; Yeo et al. 2005; Bosseno et al. 1996) and multiclonal parasite populations within individual hosts and vectors (Llewellyn et al. 2011), requiring strains to be biologically cloned prior to genotyping, a laborious caveat often overlooked by researchers.

In this study we applied high resolution nuclear and mitochondrial genotyping to contemporaneous biologically-cloned TcI strains, isolated from triatomines and mammalian hosts in Bolivia, to identify key determinants of sylvatic *T. cruzi* genetic diversification. We also explore genetic diversity and potential hybridization along two ecological clines, firstly between highland and lowland Bolivia and secondly within lowland Bolivia itself. Finally, we examine the spatial genetic structure of natural TcI populations and consider the implications of our data for human Chagas disease transmission in Bolivia.
Materials & Methods

Study area and parasite sampling

Parasite strains were isolated from sylvatic terrestrial and arboreal transmission cycles in five localities across three departments in Bolivia (Cochabamba, Potosí and Beni) (Figure 1). Study sites were situated at altitudes that ranged from ~143 to 3200 m and selected to span five major ecoregions: savannah grassland and Madeira-Tapajós moist forests (Beni), dry Andean puna and Yungas (Cochabamba) and wet Andean puna (Potosí). Parasite sampling was undertaken from 2003 - 2010 and is described for each study site individually.

Cotopachi, Cochabamba department is a densely populated area of open dry Andean puna (thorny scrub vegetation interspersed with rocky outcrops and large, spiny cacti), located ~20 km south-west of Cochabamba city at an elevation of ~2600 m. Here, parasites were sampled from wild *T. infestans* by manual microhabitat dissection and live-baited Noireau traps (Noireau et al. 2002) and from mammals (*Akodon boliviensis* and *Phyllotis ocelae*) following capture using a combination of baited Sherman and Tomohawk traps and spool-and-line tracking (Miles et al. 1981).

Triatomine sampling (*T. infestans* and *T. guasayana*) was also undertaken in neighbouring Toro Toro, an area of similar ecology to Cotopachi, situated at ~2700 m. North of Cotopachi, sylvatic *Rhodnius robustus* were collected from Chapare, a dense temperate montane forest (‘Yungas’) in the westernmost foothills of the Andes.

South of Cochabamba, parasites were isolated from wild *T. infestans* in Tupiza, a region of high altitude (~3200 m) wet Andean puna (montane grasslands) in South East Potosí department.

Sampling was undertaken in two regions of Beni department, a sparsely populated province in eastern lowland Bolivia. Ecologically Beni is a patchwork of two principal vegetation types. The majority of the department is covered by lush savannah grassland (‘Llanos de Moxos’). Along riverine alluvial plains and to the northern and western borders of the area, this ecotope is supplanted by dense Amazonian moist forests. To the east, Beni borders another moist forest (‘Madeira-Tapajós’ forest), which extends into Brazil and Santa Cruz department. In East Beni (Nueva Alianza, San Juan de Aguas Dulces, and San Juan de Mocovi), parasites were isolated from triatomines (*Rhodnius pictipes*) and mammals (*Didelphis marsupialis, Philander opossum* and *Sciureus* species) in areas of savannah grassland, interspersed with large evergreen palm trees, on the boundary between Llanos de Moxos and the moist forests of north-western Santa Cruz. The study sites in North Beni (Mercedes, San Cristobal and Santa Maria de Apere) were remote, largely uninhabited, open savannah grasslands with occasional lone standing trees, bordered by riverine forests. Here parasites were isolated from *R. robustus, P. opossum* and *D. marsupialis*. Both study sites in North and East Beni were situated at low-lying altitudes (~143 m and
~160 m, respectively) and parasite sampling was undertaken using similar methods described for other departments. All parasite strains were isolated by direct inoculation of triatomine faeces or heparinized venous animal blood into biphasic hemoculture media (Miles 1993).

Parasite strains and DTU-genotyping

A panel of 199 biological clones derived from 68 T. cruzi TcI isolates was assembled for analysis (Table S1). Biological clones were obtained from primary cultures by plate cloning according to Yeo et al. 2007 to minimize any loss of genetic diversity incurred by long-term maintenance in culture. Parasites (epimastigotes) were expanded to logarithmic phase at 28°C in RPMI-1640 liquid media supplemented with 0.5% (w/v) tryptone, 20 mM HEPES buffer (pH 7.2), 30 mM haemin, 10% (v/v) heat-inactivated fetal calf serum, 2 mM sodium glutamate, 2 mM sodium pyruvate and 25µg/ml gentamycin (all Sigma-Aldrich, UK). Genomic DNA was extracted using the Gentra PureGene Tissue Kit (Qiagen, UK), according to the manufacturer’s protocol. Clones were initially genotyped to DTU level using a triple-marker assay (Lewis et al. 2009) and classified a priori into five populations according to geographical origin: Cochabamba (n=28), Tupiza (n=15), Toro Toro (n=43), North Beni (n=26) and East Beni (n=87).

High Resolution Genotyping: Multilocus Microsatellite Typing (MLMT)

Twenty-six microsatellite loci were amplified for all 199 clones, as previously described by Llewellyn et al. 2009a. These markers are distributed across ten putative chromosomes, including six groups of physically linked loci (Weatherly et al. 2009). A full list of microsatellite targets and primers are given in Table S2. Allele sizes were determined using an automated capillary sequencer (AB3730, Applied Biosystems, UK), in conjunction with a fluorescently tagged size standard, and were manually checked for errors. All isolates were typed ‘blind’ to control for user bias.

High Resolution Genotyping: Mitochondrial Multilocus Sequence Typing (mtMLST)

Ten maxicircle gene fragments were sequenced for a subset of 78 clones, chosen to be representative of total nuclear genetic diversity (Messenger et al. 2012). For ND4 an alternate set of primers was designed to improve amplification efficiency: ND4 Fwd (5’-TTYTTCACAATGTATBGMTAG-3’) and ND4 Rvs (5’-TGTATTAYCGAYCAATTYG-3’), and reactions were performed using the same conditions as previously (Messenger et al. 2012).
Microsatellite Analysis

Individual-level sample clustering was initially defined using a Neighbour-Joining (NJ) tree based on pair-wise distances ($D_{AS}$: 1 – proportion of shared alleles at all loci/n) between microsatellite genotypes calculated in MICROSAT v1.5d (Minch et al. 1997) under the infinite-alleles model (IAM). To accommodate multi-allelic genotypes (≥3 alleles per locus), a script was written in Microsoft Visual Basic to generate random multiple diploid re-samplings of each multilocus profile (software available upon request). A final pair-wise distance matrix was derived from the mean across multiple re-sampled datasets and used to construct a NJ phylogenetic tree in PHYLIP v3.67 (Felsenstein 1989). Majority rule consensus analysis of 10,000 bootstrap trees was performed in PHYLIP v3.67 by combining 100 bootstraps generated in MICROSAT v1.5d, each drawn from 100 respective randomly re-sampled datasets.

A second analysis to define the number of putative populations in the dataset was performed using a nonparametric approach (free from Hardy-Weinberg assumptions). A K-means clustering algorithm, implemented in adegenet (Jombart et al. 2008), was used to determine the optimal number of ‘true’ populations, by reference to the Bayesian Information Criterion (BIC), which reaches a minimum when approaching the best supported assignment of individuals to the appropriate number of clusters. The relationship between these clusters and the individuals within them was then evaluated via a discriminant analysis of principal components (DAPC) according to Jombart et al. 2010.

A single randomly sampled diploid dataset was used for all subsequent analyses. Population-level genetic diversity was evaluated using sample size corrected allelic richness ($A_r$) in FSTAT 2.9.3.2 (Goudet 1995). In addition, mean $F_{IS}$, which measures the distribution of heterozygosity within and between individuals, was calculated per population in FSTAT 2.9.3.2. $F_{IS}$ can vary between -1 (all loci heterozygous for the same alleles) and +1 (all loci are homozygous for different alleles). $F_{IS} = 0$ indicates Hardy-Weinberg allele frequencies. Sample size corrected private (population-specific) allele frequency per locus (PA/L) was calculated in HP-Rare (Kalinowski 2005).

Population subdivision was estimated using pair-wise $F_{ST}$, linearised with Slatkin’s correction, in ARLEQUIN v3.11 (Excoffier et al. 2005). Statistical significance was assessed via 10,000 random permutations of alleles between populations. Within population subdivision was evaluated in ARLEQUIN v3.11 using a hierarchal Analysis of Molecular Variance (AMOVA). Population-level heterozygosity indices were also calculated in ARLEQUIN v3.11 and associated significance levels for $p$-values derived after performing a sequential Bonferroni correction to minimise the likelihood of Type 1 errors (Rice 1989). Multilocus linkage disequilibrium, estimated by the Index of Association ($I_A$) was calculated in MULTILOCUS 1.3b (Agapow & Burt 2001) and statistical significance was evaluated by comparison to a null
distribution of 1000 randomisations. Mantel’s tests for the effect of isolation by distance within populations (pair-wise genetic vs. geographic distance) were implemented in GenAIEx 6.5 using 10,000 random permutations (Peakall & Smouse 2012).

**Mitochondrial Analysis**

Sequence data from ten maxicircle gene fragments were concatenated for each isolate according to Messenger et al. 2012 and are available from GenBank under the accession numbers listed in Table S1. Additional mtMLST data from 24 previously published Tcl strains were included in selected analyses, as indicated (Messenger et al. 2012). The most appropriate nucleotide substitution model was selected from 1,624 candidates, based on the Akaike Information Criterion (AIC), in jMODELTEST 2.1.4 (Darriba et al. 2012). Alternate Maximum-Likelihood (ML) phylogenies were constructed using the TrN+G model (six substitution rate categories) in MEGA 5.10 (Tamura et al. 2011). Bootstrap support for clade topologies was estimated following the generation of 1000 pseudo-replicate datasets. Bayesian phylogenetic analysis was performed with MrBAYES, implemented through TOPALi v2.5, using the best-fit model based on the BIC (GTR+G) (Milne et al. 2009). Five independent analyses were run for one million generations, with sampling every 100 simulations (30% burn-in). Statistically-supported topological incongruence between alternate mitochondrial and nuclear phylogenies was evaluated using Kishino-Hasegawa (KH) (Kishino & Hasegawa 1989) and Shimodaira-Hasegawa (SH) (Shimodaira & Hasegawa 1999) likelihood tests in PAML v.4 (Yang 2007). Haplotype diversity (Hd) was calculated using DnaSP v5.10.1 (Librado & Rozas 2009).
Results

Strain Characteristics

One hundred and ninety-nine biological clones were genotyped across 26 polymorphic microsatellite loci (Supporting file S1). In total, 10122 alleles were identified, corresponding to 178 unique multilocus genotypes (MLGs). Multiple (≥3) alleles were observed at 0.83% of loci. Levels of intra-strain genetic diversity were high; multiclonality was observed in 65 (out of 68) uncloned strains. Identical intra-clonal genotypes were sampled in five isolates (1/18 Toro Toro, 1/11 Cochabamba and 3/26 East Beni). Clones were initially categorized into five populations based on geographical origin, consisting of three high (Cochabamba, Tupiza and Toro Toro) and two low altitude groups (North and East Beni). All populations demonstrated uniformly high numbers of unique MLGs and low frequencies of repeated MLGs (Table 1).

Nuclear genetic clustering among isolates

Patterns of isolate clustering were evaluated using two different methodologies: nonparametric population assignment (DAPC) and a NJ analysis based on pair-wise genetic distances ($D_{AS}$). Ten genetic clusters were defined among the 199 clones submitted to DAPC, once three principal components (PCs) were retained and analysed (representing 80% of the total variation). A full list of isolate assignments to DAPC populations is included in Table S1 and a multidimensional scaling plot of the DAPC results is shown in Figure 2. We observed a slight ‘elbow’ in the distribution of the BIC values across optimal cluster numbers at $K=10$ (Figure 2). DAPC-derived clusters were largely congruent with *a priori* allocations of strains to geographical populations. The ten DAPC clusters separated into three genetically distinct groups: highlands (clusters 1, 8 and 10), lowlands 1 (clusters 2, 3 and 6) and lowlands 2 (clusters 4, 5, 7 and 9). The highlands group corresponded exclusively to samples from Cochabamba, Tupiza and Toro Toro, with the exception of a single clone from *R. robustus* in Chapare (CV-05 c11), which was instead assigned to cluster 2 in the lowlands 1 group. Within the highlands group, isolates from different sampling areas and sources (hosts and vectors) were distributed across clusters 8 and 10, while cluster 1 comprised only a subset of clones from *T. infestans* found in Tupiza and Toro Toro.
The lowlands 1 group encompassed all strains from North Beni (only cluster 2) and roughly half of the isolates from *Rhodnius* spp. and *D. marsupialis* in East Beni (interspersed among clusters 2, 3 and 6). Lastly, the lowlands 2 group contained all remaining East Beni clones, including those isolated from *Rhodnius* spp., *D. marsupialis*, *P. opossum* and *Scuriurus* spp.

A NJ tree based on the same microsatellite data was constructed and further corroborated the DAPC strain assignments. A clear division between highland and lowland populations was observed, with isolates segregating into two well-supported clades (64% BS) (Figure 3). Similar to the DAPC results, the *DAS* topology supported the delineation of isolates from Beni into two groups (71% BS), one composed of all North Beni clones and the same portion of East Beni clones (*DAS lowlands 1*), the other containing the remaining East Beni strains (*DAS lowlands 2*). As previously, CV-05 cl1 from Chapare clustered as an outlier amongst North and East Beni isolates. Comparison of branch lengths in Figure 3 between the two lowland populations indicated high and consistent levels of genetic variation across strains. By contrast, highland isolates were less diverse overall (mean pair-wise *DAS* = 0.151 and 0.425 for highlands and lowlands, respectively). Within this clade, there was strong evidence for the existence of local geographic clusters in Tupiza (100% BS) and Toro Toro (73% BS), which clustered basally to the remaining highland strains.

**Population Characteristics**

Population genetic indices were calculated using both *a priori* geographical and DAPC/*DAS*-supported strain assignments (Table 1). Overall a clear division in genetic diversity and heterozygosity was apparent between highland and lowland areas. The three highland populations were characterized by lower levels of genetic diversity, as evidenced by smaller estimates of allelic richness (*A*ᵣ = 1.92 – 2.22) and numbers of private alleles per locus (PA/L = 0.19 – 0.42), compared to the lowlands (*A*ᵣ = 3.40 and 3.93 and PA/L = 1.12 and 0.60, respectively) (Table 1 and Figure 4A). All highland groups had moderately excess heterozygosity (*F*ᵢˢ = -0.241 – 0.026, 5-13.3% polymorphic loci with significant deficit in heterozygosity), whereas both lowland populations demonstrated more pronounced deviations from H-W allele frequencies (*F*ᵢˢ = 0.176 and 0.203, 63.2 and 52.3% polymorphic loci with significant deficit in heterozygosity, respectively) (Table 1). Strongly significant multilocus linkage disequilibrium was observed among all study areas (*I*ₐ = *p*<0.0001 for all populations).

**Inter-Population Gene Flow & Intra-Population Subdivision**

Estimates of subdivision (*F*ₜₜ) between *a priori* populations support a genetic demarcation between highland and lowland areas (Table 2). Little evidence for subdivision existed among the three highland study sites (*F*ₜₜ = 0.084, 0.016 and 0.079 and *p* = 0.00089, 0.0032 and 0.0001 for Cochabamba - Tupiza, Cochabamba - Toro Toro and Tupiza - Toro Toro, respectively) or between the two lowland
populations ($F_{ST} = 0.087$ and $p < 0.0001$ for North-East Beni). However, elevated $F_{ST}$ values between closest highland and lowland study sites (Cochabamba – Beni distance = ~220 km; $F_{ST} = 0.42$ and 0.35 and $p < 0.0001$ for Cochabamba – North and East Beni, respectively) indicate very limited gene flow, suggesting a powerful role for altitude and/or ecotope in structuring parasite populations. Interestingly, the extent of genetic subdivision between the most geographically distant highland populations (Cochabamba – Tupiza; distance = ~465 km) and adjacent areas of Beni (distance = ~155 km) was equivalent ($F_{ST} = 0.084$ and 0.087, respectively).

Finally, a hierarchical AMOVA was conducted, to evaluate the distribution of genetic diversity between groups of populations (highlands vs. lowlands), among populations within groups (Cochabamba, Tupiza, Toro Toro, North Beni and East Beni), and among individuals within populations. Strikingly, 23% of total genetic variation was attributed to difference between highlands and lowlands, while 4.5% and 7% were present at the population and the individuals within populations levels, respectively.

*Mitochondrial introgression across ecological clines*

For a subset of 78 clones, ten mitochondrial gene fragments (mtMLST) were sequenced and concatenated into a 3684 bp alignment. Twenty-four unique haplotypes were identified from a total of 48 variable sites (~1.3% sequence diversity). Maximum-Likelihood and Bayesian phylogenies constructed from concatenated data were not significantly different (KH test: ML tree $L = -4845.23$, Bayesian tree $L = -4848.13$ $p = 0.12$). A second ML tree was assembled using 24 additional outgroup sequences representing known TcI mitochondrial diversity, including a small population of domestic Bolivian isolates (ANDES Bol/Chile, previously described in Llewellyn et al. 2009a and Messenger et al. 2012) (Figure 5).

The mitochondrial topology demonstrated the presence of considerable genetic variation among Bolivian TcI clones. The deepest and most robust internal branch (87/1.0) separated highland and lowland populations into two major clades, each with strongly supported internal structuring. The highland group was largely homogeneous, with a number of geographically dispersed strains sharing identical mitochondrial haplotypes. The mitochondrial topology also confirmed the existence of Tupiza and Toro Toro-specific populations (98/1.0), in agreement with the nuclear tree.

Human isolates from Cochabamba (ANDES Bol/Chile), while genetically distinct from sylvatic strains circulating in the same area (63/1.0), were grouped within the main highlands clade. As previously, lowland strains were subdivided into two well-supported clades (74/1.0) with higher overall levels of genetic diversity, compared to highland isolates ($Hd = 0.81$ and 0.84 vs. 0.54, respectively; Table 1).
While the gross topology of the mitochondrial tree was broadly concordant with that of the nuclear phylogeny, internal branch patterns were significantly incongruent (SH test: ML tree $L = -4845.86$, Bayesian tree $L = -4849.55$ and $D_{AS}$ tree $L = -5006.48$, $p = 0.001$). No evidence of recombination between highland and lowland strains was observed, even in Chapare, a zone of ecological transition. Across the more ‘gentle’ ecological cline of East-North Beni, several instances of genetic hybridization were apparent. Three clones from the mixed East-North Beni group ($D_{AS}$ lowlands 1) and three isolates from the East Beni-specific population ($D_{AS}$ lowlands 2) received unambiguously different phylogenetic positions in the maxicircle topology, and are likely the progeny of multiple, independent mitochondrial introgression events (Figure 5).

Geographical dispersal within populations

To determine the extent of spatial genetic structure (or isolation by distance (IBD)) among highland and lowland isolates, Mantel’s tests were conducted using alternate nuclear and mitochondrial datasets. Nuclear IBD was detected within both highland and lowland populations (highland $R_{XY} = 0.307$, $p<0.001$ and lowland $R_{XY} = 0.209$, $p<0.001$). However, the strength of the effect was significantly larger among lowland isolates (highland slope = $0.0002 \pm 0.00000873$; lowland slope = $0.0003 \pm 0.0000179$). Furthermore, when focusing on highland clones from approximately the same spatial scale as their lowlands counterparts (i.e. omitting the local subpopulation of Tupiza isolates identified in the $D_{AS}$ tree (n=6)), little evidence for spatial structuring remained ($R_{XY} = 0.109$, $p = 0.085$). Concordant with estimates of $F_{ST}$ between populations, the differing extent of spatial genetic structuring suggests accelerated parasite dispersal among geographically disparate highland areas by comparison to adjacent lowland foci (Figure 4B).

Interestingly, no IBD was detected in either highland ($R_{XY} = 0.068$, $p = 0.161$; slope = $0.000001 \pm 0.00000345$) or lowland ($R_{XY} = 0.119$, $p = 0.0654$; slope = $0.000001 \pm 0.00000349$) populations using mitochondrial sequence data, potentially the result of lower population genetic resolution at these loci, but also consistent with the occurrence of mitochondrial introgression among lowland isolates.
Discussion

This study exploited rigorous population genetic analyses of contemporaneous parasite clones. Herein we provide several insights into the biogeographical basis of *T. cruzi* genetic diversification in Bolivia. Additionally our study undertook an in-depth dissection of TcI spatial genetic diversity and hybridization across two ecological clines.

Lowland arboreal and highland terrestrial sylvatic populations show different genetic structures

A clear dichotomy in population structure emerged between highland and lowland areas. Lowland parasites from two adjacent arboreal transmission cycles were strongly subdivided within a restricted contact zone in East Beni (~15 km²). Deep internal nuclear branching patterns in both lowland groups were indicative of stable, undisturbed, long-term genetic diversification, with correspondingly high levels of diversity. Mitochondrial introgression occurring among genetically distinct strains in Beni support prolonged historical interactions between these two populations. Consistent with high intra-host and vector clonal diversity, these data support intense, local transmission and/or low rate of genotypic extinction (Criscone & Blouin 2006). MLGs were rarely repeated, indicating only a fraction of total population genetic diversity was sampled.

In contrast, highland populations were considerably less diverse compared to their lowland counterparts. Substantial admixture and widespread dispersal of genetically homogeneous strains was observed across more geographically disparate terrestrial highland populations, supported by little evidence of genetic sub-structuring (low $F_{ST}$). Dissimilar heterozygosity estimates between highland (excess) and lowlands (deficit) suggest a recent hybrid origin for some highland strains or fundamental differences in mating systems between these two populations (Ramirez & Llewellyn 2014). Importantly, human isolates from Cochabamba were closely related to adjacent sylvatic highland strains.
Gross differences between highland and lowland population structures may be partially explained in the context of their respective ecological niches. Most lowland parasites were isolated from Didelphimorphia mammals, prominent disease reservoirs which are susceptible to high circulating parasitaemia (Legey et al. 2003) and have a propensity for non-vectoral routes of infection, including oral transmission via predation of infected vectors or mammals (Jansen & Roque 2010; Rocha et al. 2013) and exposure to contaminated anal scent gland secretions (Carreira et al. 2001). These biological features may predispose these hosts to multiplicity of infection which will be directly related to intensity and efficiency of parasite transmission and duration and course of disease (Roellig et al. 2010; Nouvellet et al. 2013). The high levels of genetic diversity among Bolivian lowland strains are consistent with this hypothesis. While minimal parasite interaction was observed between neighbouring terrestrial and arboreal transmission cycles (high $F_{ST}$ values between Cochabamba and Beni), a single clone (CV-05 cl1) isolated from $R$. robustus in the Andean foothills, was more closely related to lowland Beni strains on the basis of both nuclear and mitochondrial markers, suggesting the existence of an additional, under-sampled transmission cycle and potential hybridization zone in Chapare, northern Cochabamba.

The remaining lowland strains were isolated from Rhodnius vectors ($R$. robustus and $R$. pictipes). In general, sylvatic Rhodnius species are promiscuous feeders, which can actively migrate at night to colonize domestic environments (Fitzpatrick et al. 2008; Feliciangeli et al. 2007), thus promoting the accumulation of mixed DTU infections (Bosseno et al. 1996; Yeo et al. 2005), as well as infra-host multiclonality and co-infections with other trypanosome species, such as Trypanosoma rangeli (Dias et al. 2014). The lower genetic diversity observed among highlands strains may reflect more restricted feeding preferences and limited independent dispersal of their host vector species $T$. infestans (<500 m) (Rabinovich & Himschoot 1990; Richer et al. 2007). As a more recent host of TcI, vector competency and carrying capacity of sylvatic $T$. infestans may also vary (Araújo et al. 2014), particularly in terms of bottlenecks during transmission, which can further reduce genetic diversity, as demonstrated in tsetse fly vectors of other digenetic trypanosome species (Ruepp et al. 1997; Oberle et al. 2010).

Ecological fitting is a driver of contemporary $T$. cruzi genetic diversification

No clear association of genotype by host or vector was observed among any sylvatic Bolivian TcI populations, with the exception of a small subset of co-clustering $T$. infestans clones sampled in Tupiza and Toro Toro (DAPC cluster 1 and $D_{AS}$ highlands; $n=15$). Previous $T$. cruzi studies that favoured constrained, extant co-evolutionary scenarios were likely limited by sampling bias (O’Connor et al. 2007); Didelphimorphia mammals continue to be over represented as sources of sylvatic TcI due to their aforementioned high circulating parasitaemia, which can facilitate greater...
hemoculture positivity rates and thus parasite isolation, as well as their ease of capture.

With improved and more exhaustive sampling strategies, Tcl has now been detected among a range of Mammalia (Lisboa et al. 2004; Herrera et al. 2005; Yeo et al. 2005; Herrera et al. 2008a; Herrera et al. 2008b; Rocha et al. 2013; Lima et al. 2014), cautioning the interpretation of putative host associations. Here we demonstrate that parasite genetic diversity was principally partitioned by ecotope: arboreal lowland or terrestrial highland. Limited gene flow between neighbouring arboreal and terrestrial transmission cycles and low levels of subdivision among similar ecotopes, spanning much larger geographical distances ($F_{ST}$), strongly suggest ecological host fitting is the predominant mechanism of sylvatic T. cruzi diversification (Llewellyn et al. 2009a; Llewellyn et al. 2009b). Our observations support a current model for wider trypanosome evolution where ecological host fitting has been proposed to define major parasite clades (Hamilton et al. 2007; Lukes et al. 2014).

Mitochondrial introgression is a common phenomenon among natural T. cruzi populations

The majority of field evidence indicates T. cruzi does not conform to strict clonality or panmixia and that recombination is common, non-obligatory and idiosyncratic, potentially involving independent exchange of kinetoplastid and nuclear genetic material and both canonical meiotic and parasexual mechanisms (Carrasco et al. 1996; Machado & Ayala 2001; Ramirez et al. 2012; Ocaña-Mayorga et al. 2010; Messenger et al. 2012; Lewis et al. 2011; Roellig et al. 2013; Baptista et al. 2014). The relative contributions of alternate mating strategies to T. cruzi population structures are as yet unclear and strongly debated (Ramirez & Llewellyn 2014; Tibayrenc & Ayala 2012; Tibayrenc & Ayala 2013).

One aim of our study was to evaluate the extent of genetic recombination within two putative hybrid zones. Due to limited sample size (only a single isolate could be recovered from the politically unstable Chapare region), we were unable to detect hybridization across the highland – lowland cline. However, mitochondrial introgression was observed among a subset of lowland strains between East and North Beni. Evidence of intra-TcI genetic exchange in a primary Amazonian forest (Carrasco et al. 1996), between domestic/peri-domestic populations in Ecuador (Ocaña-Mayorga et al. 2010) and within an endemic focus in Colombia (Ramirez et al. 2012) suggests that intensive local sampling of transmission cycles is most likely to reveal recombination.

Arboreal lowland populations in Beni provide an example of an undisturbed epidemiological situation where genetic exchange might be expected (Carrasco et al. 1996). Two divergent Tcl populations overlap in this region, one sharing affinities to Tcl populations from the Chaco region to the South (East Beni), the other with...
affiliations to Amazonian TcI to the North (North Beni) (Llewellyn et al. 2009, Lima et al. 2014). Experimental recombination in T. cruzi was shown to arise in mammalian cell cultures (Gaunt et al. 2003). The aforementioned Didelphimorphia maintain high levels of multiclonal parasite populations, providing ample opportunities for hybridization to occur. Multiple mitochondrial introgression events were detected in East Beni, which appeared independent of parasite nuclear genotype, mammalian host species and study site. Consistent with previous studies, no evidence of reciprocal nuclear hybridization was detected among recombinant strains (Messenger et al. 2012; Ramírez et al. 2012; Baptista et al. 2014). While the biological cues that initiate genetic exchange remain unresolved (Gaunt et al. 2003; Lewis et al. 2010), in these populations we speculate that asymmetric introgression may act as a mechanism to facilitate ecological fitting (e.g. host range extension or resource tracking), considering the crucial role that mitochondria play in parasite metabolism, growth and development and their elevated need to escape Muller’s ratchet compared to the nuclear genome (Neiman & Taylor 2009; Ramírez & Llewellyn 2014).

Dispersal of Chagas disease in highland Bolivia

Multiple lines of evidence suggest that there is no ‘bona fide’ sylvatic transmission cycle in the Bolivian highlands. Little spatial differentiation was detected among geographically-disparate highland populations (~465 km) and this level was comparable to that observed between neighbouring lowland areas (~155 km). Terrestrial clones also displayed limited genetic IBD, a lack of private alleles and excess heterozygosity, all potentially attributable to a recent population bottleneck and/or founder event followed by clonal propagation.

This putative accelerated parasite dispersal between highland sites in comparison with lowland areas does not accord with the ecology expected for local established sylvatic transmission. Indeed, Didelphid marsupials and Rhodnius vectors have a far greater capacity for auto-dissemination than T. infestans and smaller rodents (Richer et al. 2007). One adaptive explanation for this lack of genetic population sub-structuring and geographical isolation between highland areas, is that sustained, conserved heterozygosity provides stable versatility to facilitate survival in high altitude, fluctuating climatic conditions (Widmer et al. 1987).

Alternatively, dispersal across the highlands may be recent and anthroponotic. Substantial population genetic evidence indicates that T. infestans has a precedent for passive dissemination by human populations throughout history, initially during Pre-Incan times throughout the Western Andes (Schofield 1988; Bargues et al. 2006; Cortez et al. 2010) and subsequently, Post-Colombian, eastwards into Argentina, Paraguay, Uruguay and Brazil (Panzera et al. 2004; Piccinali et al. 2009). Trafficking of genetically homogeneous, human-infective (at least in Cochabamba), highland TcI clones is reminiscent of the epidemic propagation of hybrid T. cruzi lineages TcV and TcVI by domestic T. infestans across the Southern Cone (Lewis et al. 2011). All
highland study sites coincided with major, densely-populated, transport routes 
transecting the department of Cochabamba and Potosí and the distribution of highland 
strains closely reflected human migratory movements.

Genetic continuity between human and sylvatic strains in the highlands adjacent to 
Cochabamba by mitochondrial MLST, confirms the existence of gene flow from local 
sylvatic to domestic transmission cycles. More widespread highland domestic 
infestation with *T. infestans* might be expected if sylvatic distribution is 
anthropogenically propagated. Thus the extent to which humans are responsible for 
long-range parasite distribution throughout highland Bolivia remains to be resolved.

Importantly, the widespread dispersal of limited diversity genotypes in Bolivia has 
significant biological and medical implications with respect to virulence, 
transmissibility and drug susceptibility, and the potential risk for emergent epizootic 
Chagas disease.

**Figures**

**Figure 1.** Map of Bolivia showing distribution of sylvatic *TcI* isolates among 
different ecotopes.

Parasite strains were isolated from terrestrial and arboreal transmission cycles in five 
localities across three departments: Cochabamba, Potosí and Beni. Study sites were 
situated at altitudes that ranged from ~143 to 3200 m and spanned five different 
ecoregions: savannah grassland and Madeira-Tapajós moist forests (Beni), dry 
Andean puna and Yungas (Cochabamba) and wet Andean puna (Potosí). Origins of 
individual strains are shown by closed red circles. Circle areas are proportionate to 
sampling density. Open white circles designate five *a priori* populations: 
Cochabamba, Tupiza, Toro Toro, North Beni and East Beni used for population 
 genetic analyses. Population and department names are indicated in uppercase and 
 lowercase, respectively.

**Figure 2.** Nuclear genetic clustering among 199 sylvatic Bolivian *TcI* clones.

Multidimensional scaling plot based on DAPC analysis for 10 clusters defined via *K*- 
means clustering algorithm (10⁹ iterations, 3 PCs representing 80% of total variation 
in the dataset). BIC curve is inserted with error bars representing the standard 
development about the mean of five independent runs. Inertia ellipses correspond to the 
optimal (as defined by the BIC minimum) number of population clusters among the 
genotypes analysed. Individual clones are indicated by dots. The ten DAPC clusters 
are separated into three genetically distinct groups: highlands (clusters 1, 8 and 10), 
lowlands 1 (clusters 2, 3 and 6) and lowlands 2 (clusters 4, 5, 7 and 9).
Figure 3. Unrooted Neighbour-Joining tree based on \( D_{AS} \) values between MLGs generated from 199 sylvatic Bolivian TcI clones. \( D_{AS} \) values were calculated as the mean across 1000 random diploid re-samplings of the dataset. Branch colours indicate isolate \textit{a priori} population (Cochabamba, Tupiza, Toro Toro, East Beni and North Beni; see legend). Closed grey triangles are adjacent to nodes that receive $>$60\% bootstrap support. Isolates are grouped into three statistically-supported clades (highlands, lowlands 1 and lowlands 2). Orange stars denote clones which have phylogenetically incongruent positions between nuclear and mitochondrial topologies.

Figure 4A: Allelic richness (\( A_r \)) per microsatellite locus for grouped \textit{a priori} geographical highland (diamonds) and lowland (squares) populations.

Highland populations were characterized by smaller estimates of allelic richness (\( A_r \)), compared to the lowlands (average of \( A_r = 1.92 \pm 2.22 \) and \( 3.40 \) and \( 3.93 \), respectively). Error bars represent $\pm$ standard error about the mean. Values without error bars correspond to markers containing only a single variable locus.

B: Nuclear spatial genetic analysis among \textit{T. cruzi} isolates from highland (open circles) and lowland (closed circles) populations.

Nuclear genetic isolation by distance (IBD) was observed among lowland populations (\( R_{XY} = 0.209, p<0.001; \) slope = 0.0003 $\pm$ 0.0000179), while no spatial structure was evident among highland populations spanning a much greater geographical area (\( R_{XY} = 0.109, p=0.085; \) slope = 0.0002 $\pm$ 0.0000307).

Figure 5. Maximum-Likelihood tree constructed from concatenated maxicircle sequences for 78 sylvatic Bolivian TcI clones and 24 additional TcI isolates from across the Americas.

A Maximum-Likelihood topology was constructed from concatenated maxicircle sequences for 78 sylvatic Bolivian TcI clones and rooted using 24 additional TcI strains belonging to six previously characterised populations (\textit{AM}\textsubscript{North/Cen}, \textit{ANDES}\textsubscript{Bol/Chile}, \textit{ARG}\textsubscript{North}, \textit{BRAZ}\textsubscript{North-East}, \textit{VEN}\textsubscript{dom}, and \textit{VEN}\textsubscript{silv} from Messenger \textit{et al.} 2012). The most appropriate nucleotide substitution model was TrN+G (six substitution rate categories) based on the AIC. Branch colours indicate sample \textit{a priori} population (Cochabamba, Tupiza, Toro Toro, East Beni and North Beni; see legend). Statistical support for major clades are given as equivalent bootstraps and posterior probabilities from consensus Maximum-Likelihood (1000 pseudo-replicates) and Bayesian trees (based on the GTR+G model), respectively. Orange stars denote clones which have statistically-supported phylogenetically incongruent positions between nuclear and mitochondrial topologies.
Acknowledgements

This research was supported by the Wellcome Trust and the European Commission Framework Programme Project “Comparative epidemiology of genetic lineages of Trypanosoma cruzi” ChagasEpiNet (contract #223034). LAM was supported by a BBSRC Doctoral Training Grant and a travelling fellowship from the Chadwick Trust, University College London. The funders had no role in study design, data collection and analysis, decision to publish or preparation of the manuscript.

Author Contributions

LAM designed and performed the experiments, analysed the data and drafted the manuscript. LG participated in fieldwork, contributed materials and analysed the data. MV contributed materials and analysed the data. CH, MB, MT participated in fieldwork. FT contributed materials. MAM drafted the manuscript. MSL designed the study, participated in fieldwork, analysed the data and drafted the manuscript.

References


Araújo CAC, Waniek PJ, Jansen AM (2014) Tcl/TcII co-infection can enhance Trypanosoma cruzi growth in Rhodnius prolixus. Parasites and Vectors, 7, 94.


Barnabé C, De Meeus T, Noireau F et al. (2011) Trypanosoma cruzi discrete typing units (DTUs): microsatellite loci and population genetics of DTUs TcV and TcI in Bolivia and Peru. *Infection, Genetics and Evolution*, 11, 1752-1760.


Herrera HM, Lisboa CV, Pinho AP et al. (2008a) The coati (Nasua nasua, Carnivora, Procyonidae) as a reservoir host for the main lineages of Trypanosoma cruzi in the Pantanal region, Brazil. Transactions of the Royal Society of Tropical Medicine and Hygiene, 102, 1133-1139.

Herrera HM, Abreu UG, Keuroghlian A et al. (2008b) The role played by sympatric collared peccary (Tayassu tajacu), white-lipped peccary (Tayassu pecari), and feral pig (Sus scrofa) as maintenance hosts for Trypanosoma evansi and Trypanosoma cruzi in a sylvatic area of Brazil. Parasitology Research, 103, 619-624.


Jannin J, Salvatella R (2006) Estimacion cuantitiva de la enfermedad de Chagas en las...


Lewis MD, Ma J, Yeo M et al. (2009) Genotyping of Trypanosoma cruzi: systematic selection of assays allowing rapid and accurate discrimination of all known lineages. The American Journal of Tropical Medicine and Hygiene, 81, 1041-1049.


Llewellyn MS, Miles MA, Carrasco HJ et al. (2009a) Genome-scale multilocus microsatellite typing of *Trypanosoma cruzi* discrete typing unit I reveals phylogeographic structure and specific genotypes linked to human infection. *PLoS Pathogens, 5,* e1000410.


Miles MA, De Souza AA, Povoa M (1981a) Chagas disease in the Amazon Basin III. Ecotopes of ten triatomine bug species (Hemiptera: Reduviidae) from the vicinity of Belém, Pará State, Brazil. *Journal of Medical Entomology, 18,* 264-268.


Stanford, CA: Department of Genetics, Stanford University.


Noireau F (2009) Wild Triatoma infestans, a potential threat that needs to be monitored. Memórias do Instituto Oswaldo Cruz, 104: Suppl 1, 60-64.


Ruepp S, Furger A, Kurath U et al. (1997) Survival of Trypanosoma brucei in the tsetse fly is enhanced by the expression of specific forms of procyclin. The Journal Cell Biology, 137, 1369-1379.


Yeo M, Lewis MD, Carrasco HJ et al. (2007) Resolution of multiclonal infections of *Trypanosoma cruzi* from naturally infected triatomine bugs and from experimentally infected mice by direct plating on sensitive solid medium. *International Journal for Parasitology*, 37, 111-120.


| Population | Estimated number | Estimated frequency | MLG | Average heterozygosity (Hd) | Expected heterozygosity (He) | Observed heterozygosity (Ho) | %HD | %HE | Fis ± SE | P<|F| SE | PL | PVL ± SE | PL | PVL/L ± SE | PVL/L/N | MLG | Freq. | Max. | G/N | Population |
|------------|------------------|---------------------|-----|-----------------------------|-----------------------------|----------------------------|------|------|---------|-----|------|-----|---------|-----|---------|----------|-----|------|-----|-----|-----------|
| All Highlands | 78/87 | 90.03 ± 0.05 | 52.3 | 0.48 | 0.37 | 0.16 | 0.61 ± 0.15 | 2.95 ± 0.37 | 0.26 | 0.23 | 33.3 | 19.0 | 0.84 (9/27) | 2 | 0.81 (4/7) | 2.06 | <0.001 |
| North Beni (Highlands) | 25/28 | 63.2 | 10.5 | 4.5 | 3.79 | 0.69 | 0.42 ± 0.12 | 2.22 ± 0.20 | 0.29 | 0.24 | 30.5 | 0.26 | 0.46 (4/26) | 2 | 0.46 (4/26) | 2.56 | <0.001 |
| Tupiza (Highlands) | 14/15 | 22.2 | 11.1 | 2.2 | 2.2 | 0.21 | 0.21 ± 0.07 | 2.21 ± 0.29 | 0.23 | 0.28 | 11.1 | 1.1 | 0.41 ± 0.07 | 0.026 | 0.08 | 3.54 | <0.001 |
| Toro Toro (Highlands) | 39/43 | 20.8 | 6.7 | 2.8 | 2.89 | 0.29 | 0.19 ± 0.06 | 1.92 ± 0.21 | 0.25 | 0.20 | 22.2 | 1.1 | 0.24 ± 0.09 | 1.48 | <0.001 |
| North Beni (Lowlands) | 22/26 | 90.06 ± 0.10 | 3.54 | 0.06 | 0.36 | 0.30 | 0.60 ± 0.16 | 3.93 ± 0.39 | 0.37 | 0.45 | 10.5 | 6.32 | 0.46 (8/25) | 3 | 0.46 (8/25) | 2.70 | <0.001 |
| East Beni (Lowlands) | 78/87 | 90.06 ± 0.10 | 3.54 | 0.06 | 0.36 | 0.30 | 1.12 ± 0.29 | 3.40 ± 0.46 | 0.39 | 0.48 | 9.5 | 5.23 | 0.203 ± 0.05 | 2 | 0.203 ± 0.05 | 2.23 | <0.001 |

Table 1. Population genetic parameters for sylvatic populations of T. cruzi TcI in Bolivia.
Population designation based on a priori geographical populations and DAPC/DS strain assignments.

\[ N = \text{number of isolates in population.} \]

\[ G = \text{number of multilocus genotypes (MLGs) per population based on geographical data of 26 loci analyzed.} \]

\[ \text{Max. Freq. of MLG} = \text{frequency of the most common MLG within the population.} \]

\[ \text{H} = \text{number of alleles per locus across all loci.} \]

\[ \text{Hd} = \text{Haplotype diversity measures the uniqueness of a particular haplotype in a given population, calculated using available mitochondrial sequence data in DnaSP v5.10.1 (Librado & Rozas 2009).} \]

\[ \text{PL} = \text{number of polymorphic loci out of 26 loci analyzed.} \]

\[ \text{A_r} = \text{Allelic richness as a mean over loci ± standard error, calculated in FSTAT 2.9.3 (Goudet 1995).} \]

\[ \text{PA/L} = \text{mean number of private alleles per locus ± standard error, calculated in HP-Rare (Kalinowski 2005).} \]

\[ \text{Ho} = \text{mean observed heterozygosity across all loci.} \]

\[ \text{He} = \text{mean expected heterozygosity across all loci.} \]

\[ \%\text{HE = proportion of loci showing a significant excess in heterozygosity after a sequential Bonferroni correction (Pice 1989).} \]

\[ \%\text{HD = proportion of loci showing a significant deficit in heterozygosity after a sequential Bonferroni correction (Pice 1989).} \]

\[ \text{Ho} = \text{mean observed heterozygosity across all loci.} \]

\[ \text{He} = \text{mean expected heterozygosity across all loci.} \]

\[ \%\text{HE = proportion of loci showing a significant excess in heterozygosity after a sequential Bonferroni correction (Pice 1989).} \]

\[ \%\text{HD = proportion of loci showing a significant deficit in heterozygosity after a sequential Bonferroni correction (Pice 1989).} \]
Table 2. *F*ST values in a five way comparison between populations (p-value indicated in brackets).

<table>
<thead>
<tr>
<th>Population</th>
<th>Cochabamba (Highlands)</th>
<th>Tupiza (Highlands)</th>
<th>Toro Toro (Highlands)</th>
<th>North Beni (Lowlands)</th>
<th>East Beni (Lowlands)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cochabamba (Highlands)</td>
<td>0.087 (0.000 +/- 0.000)</td>
<td>0.26 (0.000 +/- 0.000)</td>
<td>0.40 (0.000 +/- 0.000)</td>
<td>0.42 (0.000 +/- 0.000)</td>
<td>0.35 (0.000 +/- 0.000)</td>
</tr>
<tr>
<td>Tupiza (Highlands)</td>
<td>0.016 (0.003 +/- 0.000)</td>
<td>0.35 (0.000 +/- 0.000)</td>
<td>0.40 (0.000 +/- 0.000)</td>
<td>0.25 (0.000 +/- 0.000)</td>
<td>0.50 (0.000 +/- 0.000)</td>
</tr>
<tr>
<td>Toro Toro (Highlands)</td>
<td>0.79 (0.000 +/- 0.000)</td>
<td>0.26 (0.000 +/- 0.000)</td>
<td>0.40 (0.000 +/- 0.000)</td>
<td>0.50 (0.000 +/- 0.000)</td>
<td>0.087 (0.000 +/- 0.000)</td>
</tr>
<tr>
<td>North Beni (Lowlands)</td>
<td>0.42 (0.000 +/- 0.000)</td>
<td>0.25 (0.000 +/- 0.000)</td>
<td>0.50 (0.000 +/- 0.000)</td>
<td>0.087 (0.000 +/- 0.000)</td>
<td>0.35 (0.000 +/- 0.000)</td>
</tr>
<tr>
<td>East Beni (Lowlands)</td>
<td>0.35 (0.000 +/- 0.000)</td>
<td>0.26 (0.000 +/- 0.000)</td>
<td>0.40 (0.000 +/- 0.000)</td>
<td>0.50 (0.000 +/- 0.000)</td>
<td>0.087 (0.000 +/- 0.000)</td>
</tr>
</tbody>
</table>

Table S1. Panel of Bolivian *T. cruzi* Cnxz TcI biological clones assembled for analysis.

Table S2. Panel of microsatellite loci and primers employed in this study.

Supplementary File S1. Microsatellite allele sizes amplified at 26 loci across 199 TcI clones.
4.3.2 Impact of ecological disturbance on T. cruzi population structure

*T. cruzi* TcI is ubiquitous in sylvatic transmission cycles throughout its endemic range, where it is eclectic with respect to ecotope, vertebrate hosts and insect vectors. Diversity of sylvatic TcI hosts across Brazil, an ecological mosaic, is highly variable, ranging from caviomorph rodents in the Caatinga scrubland (Herrera *et al.*, 2005), to lion tamarins in the tropical Atlantic forest (Lisboa *et al.*, 2004), feral pigs and other carnivores in the Pantanal wetland (Herrera *et al.*, 2008a; 2008b; Rocha *et al.*, 2013) and primates, marsupials and rodents in Amazonia (Yeo *et al.*, 2005). To explore the interaction between intra-TcI genetic heterogeneity and ecological biodiversity, high resolution nuclear (Llewellyn *et al.*, 2009a) and mitochondrial (Messenger *et al.*, 2012) genotyping was used to characterize 107 sylvatic TcI strains, isolated from five distinct Brazilian biomes: Amazonia, Cerrado, Caatinga, Pantanal and the Atlantic Forest.

This study is reported in full below in Lima *et al*. 2014.

In summary:

- Isolates were grouped into nine populations which all displayed considerable intra-TcI genetic diversity, as evidenced by measurements of allelic richness (A_r). Comparisons of A_r and the standard deviations (SD) around D_AS (mean pair-wise nuclear genetic distances), grouped populations into those which were uniformly diverse across all samples (high A_r and low SD: e.g. Goiás, Cerrado and Para, Amazonia) and others that demonstrated intra-population substructuring (high A_r and high SD: e.g. Pantanal, Atlantic forest, Tocantins, Cerrado and Piaui, Caatinga).
- A highly divergent, geographically-dispersed, homogeneous group of strains isolated from rodents (Pantanal), Didelphimorphia, *R. prolixus*, lion tamarins (all Atlantic Forest) and bats (Cerrado) was identified, which clustered outside the known genetic diversity of TcI in the Americas, but was also distinct from the novel DTU TcBat.
- Comparison of nuclear and mitochondrial topologies revealed a single incidence of mitochondrial introgression in an isolate from *D. albiventris* in Caatinga which possessed a mitochondrial haplotype of Amazonian origin.
- Nuclear clustering indicated genetic admixture was common among strains from the Atlantic Forest which can be explained by both ongoing habitat fragmentation and long-range sylvatic introductions of TcI from distant populations, likely facilitated by volant mammals. The direct correlation between human disruption and parasite population structuring, highlighted the potential to exploit measurements of *T. cruzi* genetic diversity as a proxy for overall ecosystem health.
COVER SHEET FOR EACH ‘RESEARCH PAPER’ INCLUDED IN A RESEARCH THESIS

Please be aware that one cover sheet must be completed for each ‘Research Paper’ included in a thesis.

1. For a ‘research paper’ already published

   1.1. Where was the work published? Parasites and Vectors

   1.2. When was the work published? 2014

       1.2.1. If the work was published prior to registration for your research degree, give a brief rationale for its inclusion

       N/A


1.3. Was the work subject to academic peer review? Yes

1.4. Have you retained the copyright for the work? Yes/ No

   If yes, please attach evidence of retention.

   If no, or if the work is being included in its published format, please attach evidence of permission from copyright holder (publisher or other author) to include work

2. For a ‘research paper’ prepared for publication but not yet published

   2.1. Where is the work intended to be published? 

   2.2. Please list the paper’s authors in the intended authorship order


2.3. Stage of publication – Not yet submitted / Submitted / Undergoing revision from peer reviewers’ comments / In press

3. For multi-authored work, give full details of your role in the research included in the paper and in the preparation of the paper. (Attach a further sheet if necessary)

The candidate was responsible for generating, assembling and analyzing all mitochondrial sequence data. The candidate also participated in overall data analysis and drafting of the final manuscript.

NAME IN FULL (Block Capitals) LOUISA ALEXANDRA MESSENGER

STUDENT ID NO: 223021

CANDIDATE’S SIGNATURE LOUISA ALEXANDRA MESSENGER Date 10/2/15

SUPERVISOR/SENIOR AUTHOR’S SIGNATURE (3 above)
Wild *Trypanosoma cruzi* I genetic diversity in Brazil suggests admixture and disturbance in parasite populations from the Atlantic Forest region

Valdirene S Lima, Ana M Jansen, Louisa A Messenger, Michael A Miles and Martin S Llewellyn

**Abstract**

**Background:** *Trypanosoma cruzi* (Kinetoplastida, Trypanosomatidae) infection is an ancient and widespread zoonosis distributed throughout the Americas. Ecologically, Brazil comprises several distinct biomes: Amazonia, Cerrado, Caatinga, Pantanal and the Atlantic Forest. Sylvatic *T. cruzi* transmission is known to occur throughout these biomes, with multiple hosts and vectors involved. Parasite species-level genetic diversity can be a useful marker for ecosystem health. Our aims were to: investigate sylvatic *T. cruzi* genetic diversity across different biomes, detect instances of genetic exchange, and explore the possible impact of ecological disturbance on parasite diversity at an intra-species level.

**Methods:** We characterised 107 isolates of *T. cruzi* I (TcI; discrete typing unit, DTU I) from different major Brazilian biomes with twenty-seven nuclear microsatellite loci. A representative subset of biologically cloned isolates was further characterised using ten mitochondrial gene loci. We compared these data generated from Brazilian TcI isolates from around America.

**Results:** Genetic diversity was remarkably high, including one divergent cluster that branched outside the known genetic diversity of TcI in the Americas. We detected evidence for mitochondrial introgression and genetic exchange between the eastern Amazon and Caatinga. Finally, we found strong signatures of admixture among isolates from the Atlantic Forest region by comparison to parasites from other study sites.

**Conclusions:** Atlantic Forest sylvatic TcI populations are highly fragmented and admixed by comparison to others around Brazil. We speculate on: the possible causes of Atlantic Forest admixture; the role of *T. cruzi* as a sentinel for ecosystem health, and the impact disrupted sylvatic transmission cycles might have on accurate source attribution in oral outbreaks.

---

**Background**

*Trypanosoma cruzi* (Kinetoplastida, Trypanosomatidae) infection is an ancient and widespread zoonosis distributed throughout the Americas south of 33° latitude, where it infects approximately 8 million people [1,2]. *T. cruzi* is eclectic in terms of its mammalian hosts and haematophagous triatomine vectors. Several hundred species of mammal and many of the 140 extant triatomine species maintain transmission of *T. cruzi* in wild (sylvatic) transmission cycles [2-4]. Transmission to the host occurs usually via contamination of the mucosae or abraded skin with infected vector faeces. Oral transmission to humans via contaminated foods, especially fruit juices and sugar cane, is increasingly reported, and suspected to occur widely among sylvatic mammals through opportunistic insectivory of triatomines [5].

*T. cruzi* population genetic diversity is well described at a species level. Six discrete typing units (DTUs) are now accepted by international consensus [6]. Dates for the origin of *T. cruzi* in the Americas range between 5 and 1 MYA (calibrated biogeographically at 100 MYA) [7-9]. Estimates for the MRCA of TcI strains, arguably the most widely dispersed and abundant of all the DTUs, are younger: 1.3-0.2 MYA [7]. Nonetheless, the
age of TcI in the Americas has been sufficient to see
this genotype expand throughout multiple ecological
settings, from Amazonian forests [10] to highland An-
dean puna [11]. Furthermore, the last 1.3-0.2 MYA in
Latin America have seen intense climatic fluctuations,
including at least two glaciations [12]. The impact of
Pleistocene cycles of warming and cooling on the bio-
mic, ecological and species diversity of Latin America, in
particular in Brazil and the Brazilian Amazon, are a matter
of long debate [13]. Nonetheless, there is evidence that
historical cycles of forest expansion, contraction and
fragmentation have impacted on the current ecology
of Brazil, including small mammal distribution and
diversity [14].

Today the terrestrial ecology of Brazil is summarized
by several distinct biomes or ‘ecoregions’ [15]. The lar-
gest of these is the Amazon basin to the north, bordered
by the dryer Cerrado and seasonally flooded Pantanal to
the south. North-eastern Brazil is dominated by the xeric
scrubland of the Caatinga. Along the Atlantic coast of
Brazil south of Recife, a tropical forest ecosystem, the
Atlantic Forest, predominates. The diversity of wild TcI
hosts across this ecological mosaic is striking: cavi-
omorph rodents in the Caatinga [16]; lion tamarins in the
Atlantic forest [17]; coatis, peccaries and félid carnivores
in the Pantanal [18-20]; and multiple species of pri-
mates, marsupials and rodents in Amazonia [2]. Some
important genera are widespread – especially Didelphid
opossums. Human Chagas disease was once widespread
in Brazil, especially in central and southern parts of the
country [21]. Indeed, Chagas disease has probably been
endemic in human populations in Brazil since the earli-
est human settlements more than 10,000 years ago. It is
important not to overlook the impact that humans, an
abundant and mobile T. cruzi host species, present
throughout all Brazilian ecoregions, may have had on
contemporary parasite diversity.

Parasite alpha diversity at a species level is recognised
as a marker for ecosystem persistence, productivity,
organization and resilience [22]. Put simply, those eco-
systems in which host organisms are parasitized by an
array of different parasite species, fairly evenly distrib-
uted among hosts and host species, are considered to be
healthy. Furthermore, parasites, with their short life-
cycles and rapid mutational turnover with respect to
their hosts, can facilitate fine-scale analyses of host
population dispersal and differentiation [23]. However,
close association between host and parasite species is a
prerequisite for the use of parasite genetic diversity to
track host populations. Multi-host parasite lineages like
TcI are therefore unsuitable for such applications. None-
theless, there is some evidence that habitat fragmenta-
tion impacts on both T. cruzi diversity and prevalence of
infection [24-26]. Thus, alpha diversity in a multi-host
parasite like T. cruzi might be a useful proxy for parasite
diversity as a whole, and thus for ecosystem health.

Multilocus microsatellite typing (MLMT) is now a
widely established means of defining genetic diversity
among TcI isolates and clones [27]. Simultaneous anal-
ysis of multilocus sequence data from the mitochondrial
(maxicircle) genome (mMLST) provides a proven means
of detecting genetic exchange among clones [25,28].
Here we undertook a comparison of representative TcI
isolates from across the ecological diversity of Brazil,
examining the relationship between biomes and diversity
within biomes. We found considerable genetic diversity
among several populations, and multiple instances of
gene admixture, especially in the Atlantic Forest re-

gion. We consider these data, and the potential affect of
human-mediated habitat fragmentation on the diversity
of wild TcI in Brazil.

Methods

Parasite strains and biological cloning

One hundred and seven strains, the great majority sam-
ples from mammalian reservoir hosts captured at syl-
vatic foci throughout Brazil, were assembled for analysis
and their genotype confirmed as TcI via sequencing of a
short fragment of the glucose-6-phosphate isomerase
(GPI) gene [29]. Details of strain origin are given in
Additional file 1: Table S1 and geographic distribution in
Figure 1. A total of fourteen strains were selected from
across all biomes and biologically cloned using the plate
cloning technique described by Yeo et al. [30].

Microsatellite analysis

Twenty-seven microsatellite loci, distributed across eight
putative chromosomes, were amplified following previ-
ously described protocols across 107 strains [27]. A re-
duced subset of 19 microsatellites was employed to
evaluate diversity among a larger panel of 161 samples
including the original strains, derived clones and thirty-
three previously published multilocus microsatellite pro-
files [28]. Population genetic diversity parameters were
first calculated from sample groupings based on geog-
raphy and biome for the full 27 locus dataset (Table 1).
There were nine such groupings, as identified in Figure 1
and listed in Additional file 1: Table S1. Population-level
genetic diversity was assessed first using sample size cor-
corrected allelic richness (Ar) in FSTAT 2.9.3.2 [31]. Sec-
ondly, to provide a better measure on intra-population
sub-clustering, mean pairwise DAS and associated stand-
ard deviation was also evaluated per population. FIS, a
measure of the distribution of heterozygosity within and
between individuals, was estimated per locus per popula-
tion in FSTAT 2.9.3.2 [31]. Tests for population specific
departures from Hardy Weinberg Equilibrium at specific
loci were calculated in ARLEQUIN v3.1 and associated

Figure 1 Composite map and multidimensional scaling plot depicting sample clustering by biome and geography among 107 Trypanosoma cruzi I isolates.

Table 1 Population genetic parameters across nine Trypanosoma cruzi I populations sampled from five biomes in Brazil

<table>
<thead>
<tr>
<th>Population</th>
<th>N</th>
<th>( A_r \pm SE )</th>
<th>( D_{AS} \pm SD )</th>
<th>% PL ( H^e )</th>
<th>% PL ( H^d )</th>
<th>( F_{IS} \pm SE^c )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ceara</td>
<td>14</td>
<td>1.746 ± 0.121</td>
<td>0.290 ± 0.131</td>
<td>0</td>
<td>0</td>
<td>0.020 ± 0.012</td>
</tr>
<tr>
<td>Goais</td>
<td>4</td>
<td>1.734 ± 0.101</td>
<td>0.136 ± 0.067</td>
<td>0</td>
<td>0</td>
<td>-0.526 ± 0.032</td>
</tr>
<tr>
<td>PARA NORTH</td>
<td>28</td>
<td>2.134 ± 0.143</td>
<td>0.445 ± 0.082</td>
<td>0</td>
<td>19.2</td>
<td>0.147 ± 0.008</td>
</tr>
<tr>
<td>PARASOUTH</td>
<td>5</td>
<td>2.027 ± 0.152</td>
<td>0.416 ± 0.053</td>
<td>0</td>
<td>0</td>
<td>0.250 ± 0.019</td>
</tr>
<tr>
<td>Pantanal</td>
<td>13</td>
<td>1.698 ± 0.121</td>
<td>0.219 ± 0.197</td>
<td>26.3</td>
<td>5.2</td>
<td>0.068 ± 0.029</td>
</tr>
<tr>
<td>Piaui</td>
<td>6</td>
<td>1.930 ± 0.140</td>
<td>0.357 ± 0.188</td>
<td>0</td>
<td>0</td>
<td>0.080 ± 0.023</td>
</tr>
<tr>
<td>Atlantic Forest</td>
<td>27</td>
<td>2.010 ± 0.133</td>
<td>0.369 ± 0.199</td>
<td>33.3</td>
<td>33.3</td>
<td>0.077 ± 0.015</td>
</tr>
<tr>
<td>Santa Catarina</td>
<td>3</td>
<td>1.412 ± 0.098</td>
<td>0.057 ± 0.020</td>
<td>0</td>
<td>0</td>
<td>-0.740 ± 0.033</td>
</tr>
<tr>
<td>Tocantins</td>
<td>7</td>
<td>1.959 ± 0.133</td>
<td>0.362 ± 0.221</td>
<td>14.2</td>
<td>0</td>
<td>0.180 ± 0.025</td>
</tr>
</tbody>
</table>

\( N \) number of isolates in population.
\( A_r \) allelic richness as a mean over loci ± standard error, calculated in FSTAT.
\( D_{AS} \) mean pair-wise inverse allele sharing between samples ± standard deviation calculated in MICROSAT.
\( H^e \) proportion of loci showing significant excess heterozygosity after a sequential Bonferroni correction. Calculated in ARLEQUIN v3.1.
\( H^d \) proportion of loci showing a significant deficit in heterozygosity after a sequential Bonferroni correction. Calculated in ARLEQUIN v3.1.
\( F_{IS} \) mean FIS over loci ± standard error, calculated in FSTAT.
significance levels for p values derived after sequential Bonferroni correction to minimise the likelihood of Type 1 errors [32].

For the 19 locus dataset, individual level sample clustering was defined via a neighbour-joining tree based on pairwise distances between multilocus genotypes MLGs [evaluated using $D_{AS}$ (1 – proportion of shared alleles at all loci/n)] calculated in MICROSAT [33] (Figure 2). For the 27 locus dataset we defined genetic composition via a K-means clustering algorithm, implemented in adegenet [34], with which the optimal number of populations is defined by reference to the Bayesian Information Criterion. These groupings were subsequently submitted to a discriminant analysis of principal components (DAPC) [35], and the resulting plot is found in Figure 1.

Maxicircle analysis

Ten maxicircle sequence fragments were amplified and sequenced from fourteen $T. cruzi$ clones (see Additional file 1: Table S1 for clone identity) following previously described protocols [28]. Sequence fragments were then concatenated in each sample and aligned against previously published sequences prior to analysis [28]. Phylogenies were inferred using Maximum-Likelihood (ML)

![Figure 2](https://example.com/figure2.png)
implemented in PhyML (4 substitution rate categories) [36]. The best-fit model of nucleotide substitution was selected from 88 models and its significance evaluated according to the Akaike Information Criterion (AIC) in jMODELTEST 1.0 [37]. The best model selected for this dataset was GTR + I + G. Bootstrap support for clade topologies was estimated following the generation of 1000 pseudo-replicate datasets. Bayesian phylogenetic analysis was performed using Mr BAYES v3.1 [38] (settings according to jMODELTEST 1.0). Five independent analyses were run using a random starting tree with three heated chains and one cold chain over 10 million generations with sampling every 10 simulations (25% burn-in).

**Results**

Nuclear microsatellite loci demonstrated considerable genetic diversity among the 107 strains studied. For comparative purposes isolates were grouped *a priori* according to both geography and biome of origin (Figure 1). As such nine populations were defined. Sample assignment to these populations is presented in Additional file 1: Table S1 and population genetic parameters associated with them in Table 1. Of primary interest are sample size corrected values for allelic richness ($A_r$). $A_r$ is highest among *PARAMPARA*K and *PARASOUTH*S samples in the Eastern Amazon ($A_r = 2.027 & 2.134$), as well as in the Atlantic Forest ($A_r = 2.010$) and Tocantins, in the Cerrado ($A_r = 1.959$). While $A_r$ is a useful measure of overall sample size corrected genetic diversity, structured diversity within a population may be overlooked. We thus also calculated mean pairwise allele sharing ($D_{AS}$) between multilocus genotypes (MLGs) in each population – Table 1. The standard deviations associated with mean $D_{AS}$ values are particularly informative. Diverse populations with elevated standard deviations (e.g. Atlantic Forest – $0.369 ± 0.199$, Tocantins - $0.362 ± 0.221$) are likely to possess intra-population sub-clusters. By contrast genetic diversity is uniformly distributed among samples within populations with low standard deviations about the mean $D_{AS}$ (*PARAMPARA*K - $0.445 ± 0.082$, *PARASOUTH*S - $0.416 ± 0.053$). Observed heterozygosity varied considerably across populations. However, where population sizes ($N > 10$) are likely to facilitate meaningful interpretation, positive values for $F_{IS}$ prevailed, and by inference heterozygous deficit compared to Hardy-Weinberg expectations (Table 1).

Sample clustering based on pair-wise nuclear genetic distances provides insight into the idiosyncratic patterns of genetic diversity noted across populations. As such, considerable admixture is present between multiple populations. This phenomenon is best represented by the composite bars adjacent to the clusters in the multidimensional scaling plot displayed in Figure 1. Samples recovered from the Atlantic Forest and Tocantins cluster among multiple, divergent groups. Meanwhile TcI from *PARAMPARA*K and *PARASOUTH*S and Ceara occur among the same or closely related clusters. Remaining clusters represent intermediates between these two extremes. In summary, genetic diversity among some populations looks considerably more fragmented than among others. Mean pair-wise values for $D_{AS}$ and their associated SD seem to reflect this (Table 1).

Given the intense degree of admixture and substructure in several populations we decided not to calculate population specific linkage disequilibrium indices. Substructure is known to inflate such measures and increase the likelihood of a type 1 error [39]. Instead we chose to evaluate congruence between nuclear and mitochondrial genome clustering as evidence for rare genetic exchange events. To make such a comparison we incorporated previously published nuclear and mtDNA data into our dataset [28]. Figure 2 shows the resulting trees and the single recombinant we were able to detect among the 14 clones assayed – 6824, isolated from *Didelphis albiventris* in the Caatinga, possesses a mitochondrial genome of Amazonian origin. The hypothetical direction of the introgression event (recipient and donor) is detailed in the map inset.

The inclusion of nuclear reference microsatellite profiles from throughout the Americas in Figure 2 provides insight into the wider affinities of the Brazilian isolates. Most notably, isolates belonging to cluster 11 in Figure 1 form a homogenous group that cluster basally, well outside global TcI diversity. GPI sequences for this group nonetheless confirmed this group as TcI and no affinities with Tcbat were apparent based on the same target (data not shown).

**Discussion**

TcI diversity in Brazil is clearly considerable by comparison to that in the rest of South, Central and North America. Figure 2 shows a comparison of isolates evaluated in this study with those analysed previously [27]. Nuclear genetic data (left hand tree) indicate a clade (corresponding to population 11 in Figure 1) that lies outside the known diversity of TcI in the Americas. The presence of a bat trypanosome among this group led us to suspect that this cluster may be Tcbat, a novel DTU with affiliations to TcI originally isolated from chiroptera in Sao Paolo state, but now recognised as more widespread [40,41]. However, sequence comparison of this clade and Tcbat at the *GPI* gene rejected this hypothesis (data not shown). In contrast, all remaining TcI isolates from Brazil fall alongside their congeners, including isolates from Bolivia and Argentina, but distinct from isolates north of the Amazon basin (Venezuela, North and Central America).

The available data suggest that genetic exchange is a fairly common phenomenon among TcI isolates [25,42],
which is also capable of genetic recombination in the laboratory [43]. A consistent feature of genetic exchange events is the uniparental inheritance of mtDNA. At a population level, as well as between DTUs, these events lead to clear instances of mitochondrial introgression [25]. Thus a pair of isolates maybe highly genetically similar on a nuclear level, but lack any affinity between mitochondrial genomes. We identified one such hybrid among those clones we assayed - 6824. In a recent review, it was proposed that ‘different evolutionary pressures and molecular clocks’ between non-coding nuclear microsatellite and coding mtDNA, rather than genetic exchange, might account for such signals of introgression [44]. However, such a theory requires a situation in which two (or more) near identical nuclear genotypes (e.g. 6824 and 9667) experience radically different evolutionary pressures on their mitochondrial genomes, which end up closely resembling the mitochondrial genotype of nearby or sympatric clones, in this case from the same host (*Didelphis albiventris*). Given that this pattern of introgression fits precisely with that observed in hybrids in the laboratory [43], and between DTUs in the field (TcI/TcIV) [45], recombination is the only reasonable explanation.

Of particular interest in our study was the distribution and structure of genetic diversity within and between ecoregions. Admixture was most common in the Atlantic forest region, and largely absent from the Amazon region in Pará state (Figure 1). As such, samples from the Atlantic Forest region have strong affinity with those from around Brazil and are thus distributed across multiple genetic clusters in Figure 1. The inset in Figure 1 provides fine details of parasite genetic diversity in the Atlantic Forest region. Isolates at the northern extreme of this region have predictable affinity with samples from the Caatinga (cluster 5). However, admixture into Atlantic forest from other populations is far less predictable, especially from Amazonia, and the Pantanal, which lie thousands of kilometres from the Atlantic forest. The impact of Atlantic forest fragmentation on species abundance and diversity is well documented (e.g. [46-48]). Most studies report loss of alpha diversity correlating inversely with forest fragment size, within as well as between species [46,49]. In contrast, allelic richness indices in our study suggested substantial *T. cruzi* genetic diversity within the Atlantic Forest (Table 1). However, invasive species introductions are common in the Atlantic Forest region (e.g. [50]), and it seems that several long range introductions from distant populations may also explain the high genetic diversity of Tcl in the region. Thus, unlike Tcl populations from Amazonia and Caatinga, which generally exhibit high genetic diversity but little admixture, high genetic diversity in the Atlantic Forest region is explained by these introductions and associated admixture. Long-range sylvatic dispersal of *T. cruzi* can be achieved by bats. Indeed, the presence of *T. cruzi* clade trypanosomes in Africa can be explained by rapid aerial dispersal [51]. Cluster 11 contains several isolates from bats, which could explain the geographic diversity of isolates in this clade (Atlantic Forest, Pantanal, Cerrado), as well as its genetic homogeneity. However, other geographically diverse isolate groupings containing Atlantic Forest isolates have no link to volant mammals.

There is a circumstantial link between Atlantic Forest loss (88% of its former extent [52]), human population density, and Tcl genetic admixture in the region. *T. cruzi* infection is commonly termed a ‘zoonosis’, which implies unidirectional dispersion from sylvatic transmission cycles to man. Until the successful triatomine eradication campaigns of the 1970s and 1980s, domestic *T. cruzi* infection was endemic throughout much (although not all) of the Atlantic Forest region [21]. It is thus possible that many of these long-range introductions into the Atlantic are ‘enzooses’, i.e. Tcl strains imported via immigrant human populations, which subsequently escaped in the local sylvatic environment.

**Conclusions**

Rather like primary rainforest, ‘pristine’ sylvatic *T. cruzi* diversity may be now relatively rare in South America, especially where human population densities and infections rates have been historically high. The presence of disturbed and admixed sylvatic *T. cruzi* populations in populous areas has major implications for the effective source attribution and thus future prevention of oral outbreaks [5]. Many such outbreaks have occurred in Brazil in recent years [53]. As such, the discrimination of the source of oral outbreak strains as being from either the local wild population or from another region via the importation of foodstuffs becomes complex. This is because the local wild strains themselves may represent long-range introductions. Nonetheless, admixture among sylvatic parasite populations has a possible role as a proxy for environmental disturbance. Future approaches could involve high-resolution genotyping and focused sampling of Atlantic forest fragments, including co-variates like mammalian and insect biodiversity, to further explore the use of *T. cruzi* as a sentinel species for ecosystem health.

**Additional file**

Additional file 1: Table S1. *Trypanosoma cruzi* isolates evaluated in this study.

**Competing interests**

The authors declare they have no competing interests. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.


Atlantic Forest remnants in Southeastern Brazil.


doi:10.1186/1756-3305-7-263

Cite this article as: Lima et al.: Wild Trypanosoma cruzi I genetic diversity in Brazil suggests admixture and disturbance in parasite populations from the Atlantic Forest region. Parasites & Vectors 2014, 7:263.

Submit your next manuscript to BioMed Central and take full advantage of:

• Convenient online submission
• Thorough peer review
• No space constraints or color figure charges
• Immediate publication on acceptance
• Inclusion in PubMed, CAS, Scopus and Google Scholar
• Research which is freely available for redistribution

Submit your manuscript at www.biomedcentral.com/submit
4.3.3 Emergence of human-associated genotypes (TcI_DOM) in Central America

TcI, the most abundant and widely distributed lineage, is the principal cause of human chagasic cardiomyopathy in northern South America (Ramirez et al., 2010; Carrasco et al., 2012). Multiple molecular markers consistently identify high levels of genetic diversity within sylvatic TcI populations (Herrera et al., 2007b; 2009; O’Connor et al., 2007; Falla et al., 2009; Llewellyn et al., 2009a; Ocaña-Mayorga et al., 2010; Messenger et al., 2012; Lima et al., 2014; Messenger et al., accepted), and divergent, but genetically homogeneous, strains isolated from human infections and largely absent from mammals or vectors, henceforth TcI_DOM (previously TcIa/VENDOM) (Llewellyn et al., 2009a; Cura et al., 2010; Ramirez et al., 2012).

To investigate the evolutionary origin of TcI_DOM, high resolution nuclear (Llewellyn et al., 2009a) and mitochondrial (Messenger et al., 2012) genotyping was used to characterize 72 TcI isolates, including 22 strains from Central America (Guatemala, Mexico and Honduras) and 50 additional reference isolates, chosen to be representative of T. cruzi intra-TcI genetic diversity. The aim of this study was to determine whether TcI_DOM emerged in northern South America as a sister group of North American strains and propagated among domestic transmission cycles or whether it originated in North America, prior to dispersal into South American domestic populations.

This study is reported in full below in Zumaya-Estrada et al. 2012.

In summary:

- Nuclear and mitochondrial topologies grouped isolates into three statistically-supported populations: TcISOUTH, TcINORTH-CENT and TcIDOM. Nuclear allelic richness (A_r) and mitochondrial nucleotide diversity (π) demonstrated a hierarchical cline in genetic diversity TcISOUTH > TcINORTH-CENT > TcIDOM, consistent with phylogenetic branching patterns.
- The microsatellite phylogeny robustly grouped TcINORTH-CENT as a monophyletic clade with TcIDOM clustered firmly within it.
- Reduced genetic diversity among TcINORTH-CENT by comparison to TcISOUTH supports the hypothesis that TcI originated in South America before dispersal across the Isthmus of Panama during the Great American Interchange.
- The phylogenetic placement of TcIDOM within TcINORTH-CENT, suggests this group most likely originated in North/Central America before disseminating southwards, a finding which is consistent with the migration patterns of early colonizing Amerindians.
- The historical and current mechanisms of TcIDOM dispersal remain largely unexplained. However, its widespread geographical distribution suggests that distinct human-restricted genotypes can be sustained, despite the presence of sympatric infective sylvatic strains, due to inefficient stercorarian vector transmission.
- This study reinforces the need for additional sampling efforts from Central and South America to characterize TcIDOM further and from the Southern Cone to define the geographical distribution of this putative human-associated genotype.
Please be aware that one cover sheet must be completed for each ‘Research Paper’ included in a thesis.

1. For a ‘research paper’ already published
   1.1. Where was the work published? Parasites and Vectors
   1.2. When was the work published? 2012
   1.2.1. If the work was published prior to registration for your research degree, give a brief rationale for its inclusion
   N/A

1.3. Was the work subject to academic peer review? Yes

1.4. Have you retained the copyright for the work? Yes
   If yes, please attach evidence of retention.
   If no, or if the work is being included in its published format, please attach evidence of permission from copyright holder (publisher or other author) to include work

2. For a ‘research paper’ prepared for publication but not yet published
   2.1. Where is the work intended to be published? 

2.2. Please list the paper’s authors in the intended authorship order

2.3. Stage of publication – Not yet submitted / Submitted / Undergoing revision from peer reviewers’ comments / In press

3. For multi-authored work, give full details of your role in the research included in the paper and in the preparation of the paper. (Attach a further sheet if necessary)
   The candidate had significant input into experimental design, contributed laboratory reagents, supervision and reference datasets. The candidate also participated in mitochondrial data generation, analysis and drafting of the final manuscript.

NAME IN FULL (Block Capitals) LOUISA ALEXANDRA MESSENGER

STUDENT ID NO: 223021

CANDIDATE’S SIGNATURE

Date 10/2/15

SUPERVISOR/SENIOR AUTHOR’S SIGNATURE (3 above)
North American import? Charting the origins of an enigmatic *Trypanosoma cruzi* domestic genotype

Federico A Zumaya-Estrada, Louisa A Messenger, Teresa Lopez-Ordonez, Michael D Lewis, Carlos A Flores-Lopez, Alejandro J Martínez-Ibarra, Pamela M Pennington, Celia Cordon-Rosales, Hernan V Carrasco, Maikel Segovia, Michael A Miles and Martin S Llewellyn

**Abstract**

**Background:** *Trypanosoma cruzi*, the agent of Chagas disease, is currently recognized as a complex of six lineages or Discrete Typing Units (DTU): TcI-TcVI. Recent studies have identified a divergent group within TcI - TcIDOM. TcIDOM is associated with a significant proportion of human TcI infections in South America, largely absent from local wild mammals and vectors, yet closely related to sylvatic strains in North/Central America. Our aim was to examine hypotheses describing the origin of the TcIDOM genotype. We propose two possible scenarios: an emergence of TcIDOM in northern South America as a sister group of North American strain progenitors and dispersal among domestic transmission cycles, or an origin in North America, prior to dispersal back into South American domestic cycles. To provide further insight we undertook high resolution nuclear and mitochondrial genotyping of multiple Central American strains (from areas of México and Guatemala) and included them in an analysis with other published data.

**Findings:** Mitochondrial sequence and nuclear microsatellite data revealed a cline in genetic diversity across isolates grouped into three populations: South America, North/Central America and TcIDOM. As such, greatest diversity was observed in South America ($A_r = 4.851$, $\pi = 0.00712$) and lowest in TcIDOM ($A_r = 1.813$, $\pi = 0.00071$). Nuclear genetic clustering (genetic distance based) analyses suggest that TcIDOM is nested within the North/Central American clade.

**Conclusions:** Declining genetic diversity across the populations, and corresponding hierarchical clustering suggest that emergence of this important human genotype most likely occurred in North/Central America before moving southwards. These data are consistent with early patterns of human dispersal into South America.

**Keywords:** Trypanosoma cruzi, Maxicircle, Microsatellite, Chagas Disease, Phylogeography, Population genetics, TcI

**Findings**

*Trypanosoma cruzi*, the aetiological agent of Chagas disease, infects 6-8 million people in Latin America, while some 25 million more are at risk of acquiring the disease [1]. Parasite transmission to mammal hosts, including humans, can occur through contact with the faeces of hematophagous triatomine bugs. However, non-vectorial routes are also recognized, including blood transfusion, organ transplantation, congenital transmission, and oral transmission via ingestion of meals contaminated with infected triatomine feces [2,3].

*T. cruzi* (family Trypanosomatidae; Euglenozoa: Kinetoplastida) is most closely related to several widely dispersed species of bat trypanosomes [4]. Salivarian trypanosomes including medically important *Trypanosoma brucei* subspecies, represent a more divergent group [5]. The age of the split between the *T. cruzi*-containing and *T. brucei*-containing trypanosome lineages is thought to have been concurrent with the separation of Africa and South America/Antarctica/Australasia 100MYA [6], implying that *T. cruzi* and the other Schizotrypanum
## Table 1 Trypanosoma cruzi I samples included in this study

<table>
<thead>
<tr>
<th>Strain code</th>
<th>Strain</th>
<th>Host/vector</th>
<th>Country</th>
<th>State</th>
<th>Latitude</th>
<th>Longitude</th>
<th>Date</th>
<th>Population</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>PALDA4</td>
<td>PALDA4</td>
<td>Didelphis albiventris</td>
<td>Argentina</td>
<td>Chaco</td>
<td>-27.133</td>
<td>-61.460</td>
<td>2001</td>
<td>SOUTH</td>
<td>Messenger et al., [12]</td>
</tr>
<tr>
<td>PALDA21</td>
<td>PALDA21</td>
<td>Didelphis albiventris</td>
<td>Argentina</td>
<td>Chaco</td>
<td>-27.133</td>
<td>-61.460</td>
<td>2001</td>
<td>SOUTH</td>
<td>Messenger et al., [12]</td>
</tr>
<tr>
<td>PALDA5</td>
<td>PALDA5</td>
<td>Didelphis albiventris</td>
<td>Argentina</td>
<td>Chaco</td>
<td>-27.133</td>
<td>-61.460</td>
<td>2001</td>
<td>SOUTH</td>
<td>Messenger et al., [12]</td>
</tr>
<tr>
<td>PALDAV2</td>
<td>PALDAV2^3</td>
<td>Triatoma infestans</td>
<td>Argentina</td>
<td>Chaco</td>
<td>-27.133</td>
<td>-61.460</td>
<td>2001</td>
<td>SOUTH</td>
<td>Messenger et al., [12]</td>
</tr>
<tr>
<td>PALDA20</td>
<td>PALDA20</td>
<td>Didelphis albiventris</td>
<td>Argentina</td>
<td>Chaco</td>
<td>-27.133</td>
<td>-61.460</td>
<td>2001</td>
<td>SOUTH</td>
<td>Messenger et al., [12]</td>
</tr>
<tr>
<td>COTMA38</td>
<td>COTMA38</td>
<td>Akodon boliviensis</td>
<td>Bolivia</td>
<td>Cotopachi</td>
<td>-17.430</td>
<td>-66.270</td>
<td>2004</td>
<td>SOUTH</td>
<td>Messenger et al., [12]</td>
</tr>
<tr>
<td>SJM22</td>
<td>SJM22 c1</td>
<td>Didelphis marsupialis</td>
<td>Bolivia</td>
<td>Beni</td>
<td>-14.810</td>
<td>-64.600</td>
<td>2004</td>
<td>SOUTH</td>
<td>Messenger et al., [12]</td>
</tr>
<tr>
<td>SJM34</td>
<td>SJM34</td>
<td>Didelphis marsupialis</td>
<td>Bolivia</td>
<td>Beni</td>
<td>-14.810</td>
<td>-64.600</td>
<td>2004</td>
<td>SOUTH</td>
<td>Messenger et al., [12]</td>
</tr>
<tr>
<td>SJM37</td>
<td>SJM37</td>
<td>Didelphis marsupialis</td>
<td>Bolivia</td>
<td>Beni</td>
<td>-14.810</td>
<td>-64.600</td>
<td>2004</td>
<td>SOUTH</td>
<td>Messenger et al., [12]</td>
</tr>
<tr>
<td>SJM39</td>
<td>SJM39 c3</td>
<td>Didelphis marsupialis</td>
<td>Bolivia</td>
<td>Beni</td>
<td>-14.810</td>
<td>-64.600</td>
<td>2004</td>
<td>SOUTH</td>
<td>Messenger et al., [12]</td>
</tr>
<tr>
<td>SJM41</td>
<td>SJM41</td>
<td>Philander opossum</td>
<td>Bolivia</td>
<td>Beni</td>
<td>-14.810</td>
<td>-64.600</td>
<td>2004</td>
<td>SOUTH</td>
<td>Messenger et al., [12]</td>
</tr>
<tr>
<td>SJMC12</td>
<td>SJMC12</td>
<td>Philander opossum</td>
<td>Bolivia</td>
<td>Beni</td>
<td>-14.810</td>
<td>-64.600</td>
<td>2004</td>
<td>SOUTH</td>
<td>Messenger et al., [12]</td>
</tr>
<tr>
<td>XE5167</td>
<td>XE5167 c1</td>
<td>Didelphis marsupialis</td>
<td>Brasil</td>
<td>Para</td>
<td>-1.710</td>
<td>-48.880</td>
<td>1999</td>
<td>SOUTH</td>
<td>Messenger et al., [12]</td>
</tr>
<tr>
<td>IM4810</td>
<td>IM4810</td>
<td>Didelphis marsupialis</td>
<td>Brasil</td>
<td>Manaus</td>
<td>-3.070</td>
<td>-60.160</td>
<td>2002</td>
<td>SOUTH</td>
<td>Messenger et al., [12]</td>
</tr>
<tr>
<td>B2085</td>
<td>B2085</td>
<td>Didelphis marsupialis</td>
<td>Brasil</td>
<td>Belem</td>
<td>-1.360</td>
<td>-48.360</td>
<td>1991</td>
<td>SOUTH</td>
<td>Messenger et al., [12]</td>
</tr>
<tr>
<td>AAA1c5</td>
<td>AAA1c5</td>
<td>Rhodnius prolixus</td>
<td>Colombia</td>
<td>Casanare</td>
<td>4.150</td>
<td>-71.200</td>
<td>2010</td>
<td>SOUTH</td>
<td>Ramirez et al., Molecular Ecology In press</td>
</tr>
<tr>
<td>AAA7c2</td>
<td>AAA7c2</td>
<td>Rhodnius prolixus</td>
<td>Colombia</td>
<td>Casanare</td>
<td>5.100</td>
<td>-71.600</td>
<td>2010</td>
<td>SOUTH</td>
<td>Ramirez et al., Molecular Ecology In press</td>
</tr>
<tr>
<td>AAB3c13</td>
<td>AAB3c13</td>
<td>Rhodnius prolixus</td>
<td>Colombia</td>
<td>Casanare</td>
<td>4.150</td>
<td>-71.200</td>
<td>2010</td>
<td>SOUTH</td>
<td>Ramirez et al., Molecular Ecology In press</td>
</tr>
<tr>
<td>AAC1c3</td>
<td>AAC1c3</td>
<td>Rhodnius prolixus</td>
<td>Colombia</td>
<td>Casanare</td>
<td>5.100</td>
<td>-71.600</td>
<td>2010</td>
<td>SOUTH</td>
<td>Ramirez et al., Molecular Ecology In press</td>
</tr>
<tr>
<td>AACf1c4</td>
<td>AACf1c4</td>
<td>Canis familiaris</td>
<td>Colombia</td>
<td>Casanare</td>
<td>5.100</td>
<td>-71.600</td>
<td>2010</td>
<td>SOUTH</td>
<td>Ramirez et al., Molecular Ecology In press</td>
</tr>
<tr>
<td>AAD6c6</td>
<td>AAD6c6</td>
<td>Rhodnius prolixus</td>
<td>Colombia</td>
<td>Casanare</td>
<td>5.100</td>
<td>-71.600</td>
<td>2010</td>
<td>SOUTH</td>
<td>Ramirez et al., Molecular Ecology In press</td>
</tr>
<tr>
<td>CACQc17</td>
<td>CACQc17</td>
<td>Homo sapiens</td>
<td>Colombia</td>
<td>Santander</td>
<td>6.963</td>
<td>-73.420</td>
<td>2009</td>
<td>TcIDOM</td>
<td>Ramirez et al., Molecular Ecology In press</td>
</tr>
<tr>
<td>CACQc18</td>
<td>CACQc18</td>
<td>Homo sapiens</td>
<td>Colombia</td>
<td>Santander</td>
<td>6.644</td>
<td>-73.654</td>
<td>2009</td>
<td>TcIDOM</td>
<td>Ramirez et al., Molecular Ecology In press</td>
</tr>
<tr>
<td>DYRc116</td>
<td>DYRc116</td>
<td>Homo sapiens</td>
<td>Colombia</td>
<td>Boyacá</td>
<td>5.640</td>
<td>-72.899</td>
<td>2007</td>
<td>TcIDOM</td>
<td>Ramirez et al., Molecular Ecology In press</td>
</tr>
<tr>
<td>Sample Code</td>
<td>Species</td>
<td>Species + Code</td>
<td>Location</td>
<td>Coordinates</td>
<td>Year</td>
<td>Institute</td>
<td>Author(s)</td>
<td>Journal</td>
<td>Page</td>
</tr>
<tr>
<td>-------------</td>
<td>---------</td>
<td>----------------</td>
<td>----------</td>
<td>-------------</td>
<td>------</td>
<td>-----------</td>
<td>-----------</td>
<td>---------</td>
<td>------</td>
</tr>
<tr>
<td>EBl11</td>
<td>Homo sapiens</td>
<td>EBcl11</td>
<td>Colombia</td>
<td>Boyacá</td>
<td>5.130</td>
<td>-73.119</td>
<td>2007</td>
<td>TcIDOM</td>
<td>Ramirez et al., Molecular Ecology In press</td>
</tr>
<tr>
<td>FEc110</td>
<td>Homo sapiens</td>
<td>FEc110</td>
<td>Colombia</td>
<td>Boyacá</td>
<td>5.920</td>
<td>-73.500</td>
<td>2001</td>
<td>TcIDOM</td>
<td>Ramirez et al., Molecular Ecology In press</td>
</tr>
<tr>
<td>Td3c11</td>
<td>Triatoma dimidiata</td>
<td>Td3c11</td>
<td>Colombia</td>
<td>Boyacá</td>
<td>6.270</td>
<td>-71.200</td>
<td>2000</td>
<td>TcIDOM</td>
<td>Ramirez et al., Molecular Ecology In press</td>
</tr>
<tr>
<td>X-1084c10</td>
<td>Rhodnius prolixus</td>
<td>X-1084cl10</td>
<td>Colombia</td>
<td>Boyacá</td>
<td>4.960</td>
<td>-73.630</td>
<td>2010</td>
<td>SOUTH</td>
<td>Ramirez et al., Molecular Ecology In press</td>
</tr>
<tr>
<td>X-236c9</td>
<td>Rhodnius prolixus</td>
<td>X-236cl9</td>
<td>Colombia</td>
<td>Boyacá</td>
<td>4.960</td>
<td>-73.630</td>
<td>2010</td>
<td>SOUTH</td>
<td>Ramirez et al., Molecular Ecology In press</td>
</tr>
<tr>
<td>YAS1c3</td>
<td>Alouatta spp</td>
<td>YAS1c3</td>
<td>Colombia</td>
<td>Casanare</td>
<td>5.300</td>
<td>-72.400</td>
<td>2010</td>
<td>SOUTH</td>
<td>Ramirez et al., Molecular Ecology In press</td>
</tr>
<tr>
<td>38</td>
<td>Triatoma dimidiata</td>
<td>38</td>
<td>Guatemala</td>
<td>Jutiapa</td>
<td>14.287</td>
<td>-89.844</td>
<td>2000</td>
<td>NORTH-CENT</td>
<td>This study</td>
</tr>
<tr>
<td>46</td>
<td>Triatoma dimidiata</td>
<td>46</td>
<td>Guatemala</td>
<td>Santa Rosa</td>
<td>14.177</td>
<td>-90.303</td>
<td>2001</td>
<td>NORTH-CENT</td>
<td>This study</td>
</tr>
<tr>
<td>66</td>
<td>Triatoma dimidiata</td>
<td>66</td>
<td>Guatemala</td>
<td>Jalapa</td>
<td>14.633</td>
<td>-89.89</td>
<td>2001</td>
<td>NORTH-CENT</td>
<td>This study</td>
</tr>
<tr>
<td>67</td>
<td>Triatoma dimidiata</td>
<td>67</td>
<td>Guatemala</td>
<td>Jutiapa</td>
<td>14.287</td>
<td>-89.844</td>
<td>2001</td>
<td>NORTH-CENT</td>
<td>This study</td>
</tr>
<tr>
<td>70</td>
<td>Triatoma dimidiata</td>
<td>70</td>
<td>Guatemala</td>
<td>Jutiapa</td>
<td>14.287</td>
<td>-89.844</td>
<td>2001</td>
<td>NORTH-CENT</td>
<td>This study</td>
</tr>
<tr>
<td>71</td>
<td>Triatoma dimidiata</td>
<td>71</td>
<td>Guatemala</td>
<td>Jalapa</td>
<td>14.633</td>
<td>-89.89</td>
<td>2001</td>
<td>NORTH-CENT</td>
<td>This study</td>
</tr>
<tr>
<td>83</td>
<td>Triatoma dimidiata</td>
<td>83</td>
<td>Guatemala</td>
<td>Chiquimula</td>
<td>14.768</td>
<td>-89.458</td>
<td>2002</td>
<td>NORTH-CENT</td>
<td>This study</td>
</tr>
<tr>
<td>95</td>
<td>Triatoma dimidiata</td>
<td>95</td>
<td>Guatemala</td>
<td>Chiquimula</td>
<td>14.768</td>
<td>-89.458</td>
<td>2002</td>
<td>NORTH-CENT</td>
<td>This study</td>
</tr>
<tr>
<td>100</td>
<td>Triatoma dimidiata</td>
<td>100</td>
<td>Guatemala</td>
<td>Santa Rosa</td>
<td>14.177</td>
<td>-90.303</td>
<td>2002</td>
<td>NORTH-CENT</td>
<td>This study</td>
</tr>
<tr>
<td>113</td>
<td>Triatoma dimidiata</td>
<td>113</td>
<td>Guatemala</td>
<td>Chiquimula</td>
<td>14.768</td>
<td>-89.458</td>
<td>2002</td>
<td>NORTH-CENT</td>
<td>This study</td>
</tr>
<tr>
<td>116</td>
<td>Triatoma dimidiata</td>
<td>116</td>
<td>Guatemala</td>
<td>Baja Verapaz</td>
<td>15.079</td>
<td>-90.413</td>
<td>2002</td>
<td>NORTH-CENT</td>
<td>This study</td>
</tr>
<tr>
<td>154</td>
<td>Triatoma dimidiata</td>
<td>154</td>
<td>Guatemala</td>
<td>Alta Verapaz</td>
<td>15.594</td>
<td>-90.149</td>
<td>2002</td>
<td>NORTH-CENT</td>
<td>This study</td>
</tr>
<tr>
<td>DAVIScl1</td>
<td>Triatoma dimidiata</td>
<td>DAVIS 9.90 cl1</td>
<td>Honduras</td>
<td>Tegucigalpa</td>
<td>14.080</td>
<td>-87.200</td>
<td>1983</td>
<td>NORTH-CENT</td>
<td>Messenger et al., 2012</td>
</tr>
<tr>
<td>ANITA II</td>
<td>Triatoma dimidiata</td>
<td>ANITA</td>
<td>Mexico</td>
<td>Campeche</td>
<td>19.188</td>
<td>-90.300</td>
<td>2011</td>
<td>NORTH-CENT</td>
<td>This study</td>
</tr>
<tr>
<td>CAM6</td>
<td>Triatoma dimidiata</td>
<td>CAM6</td>
<td>Mexico</td>
<td>Campeche</td>
<td>19.188</td>
<td>-90.300</td>
<td>2011</td>
<td>NORTH-CENT</td>
<td>This study</td>
</tr>
<tr>
<td>CRISTY</td>
<td>Homo sapiens</td>
<td>CRISTY</td>
<td>Mexico</td>
<td>San Luis Potosi</td>
<td>22.159</td>
<td>-100.990</td>
<td>2007</td>
<td>NORTH-CENT</td>
<td>This study</td>
</tr>
<tr>
<td>MICH1</td>
<td>Triatoma dimidiata</td>
<td>MICH</td>
<td>Mexico</td>
<td>Michoacan</td>
<td>19.567</td>
<td>-101.707</td>
<td>2011</td>
<td>NORTH-CENT</td>
<td>This study</td>
</tr>
<tr>
<td>NINOA</td>
<td>Homo sapiens</td>
<td>NINOA</td>
<td>Mexico</td>
<td>Oaxaca</td>
<td>17.054</td>
<td>-96.714</td>
<td>1994</td>
<td>NORTH-CENT</td>
<td>This study</td>
</tr>
<tr>
<td>PLI</td>
<td>Dipetalogaster maxima</td>
<td>PL</td>
<td>Mexico</td>
<td>Baja California Sur</td>
<td>26.044</td>
<td>-111.666</td>
<td>2001</td>
<td>NORTH-CENT</td>
<td>This study</td>
</tr>
<tr>
<td>QROI</td>
<td>Triatoma barberi</td>
<td>QRO</td>
<td>Mexico</td>
<td>Queretaro</td>
<td>20.594</td>
<td>-100.393</td>
<td>1986</td>
<td>NORTH-CENT</td>
<td>This study</td>
</tr>
<tr>
<td>TQI</td>
<td>Triatoma pallidipennis</td>
<td>TQ</td>
<td>Mexico</td>
<td>Morelos</td>
<td>18.953</td>
<td>-99.223</td>
<td>1991</td>
<td>NORTH-CENT</td>
<td>This study</td>
</tr>
</tbody>
</table>

Zumaya-Estrada et al. Parasites & Vectors 2012, 5:226
http://www.parasitesandvectors.com/content/5/1/226
species evolved exclusively in South America. Others propose an alternative origin of *T. cruzi* from an ancestral bat trypanosome potentially capable of long range dispersal [7]. Whilst the precise scenario for the arrival of ancestral *Schizotrypanum* lineages in South America is a matter for debate, the current continental distribution and genetic diversity of *T. cruzi* supports an origin within South America. Parasite transmission is maintained via hundreds of mammal and triatomine species in different biomes throughout South and Central America, as well as the southern states of the USA [8].

Biochemical and molecular markers support the existence of six lineages or Discrete Typing Units (DTU): TcI, - TcVI agreed by international consensus [9]. Each DTU can be loosely associated with a particular ecological and/or geographical framework [10]. TcI is ubiquitous among arboreal sylvatic foci throughout the geographic distribution of *T. cruzi* and is the major agent of human Chagas disease in northern South America. Several molecular tools now identify substantial genetic diversity within TcI [11-14]. Importantly these new approaches consistently reveal the presence of a genetically divergent and homogeneous TcI group (henceforth TcIDOM – previously TcIa/VENDOM) associated with human infections from Venezuela to Northern Argentina, and largely absent from wild mammals and vectors sampled to date [14]. The origin of this clade is unclear, although recent work supports a

<table>
<thead>
<tr>
<th>Table 1 Trypanosoma cruzi I samples included in this study (Continued)</th>
</tr>
</thead>
<tbody>
<tr>
<td>XAL1 XAL Triatoma dimidiata Mexico Veracruz 19.173 -96.133 2003 NORTH-CENT This study</td>
</tr>
<tr>
<td>9209802P 9209802P cl1 Didelphis marsupialis USA Georgia 32.430 -83.310 1992 NORTH-CENT Messenger et al., [12]</td>
</tr>
<tr>
<td>9307 93070103P cl1 Didelphis marsupialis USA Georgia 32.430 -83.310 1993 NORTH-CENT Messenger et al., [12]</td>
</tr>
<tr>
<td>ARMA USAARMA cl3 Dasypus novemcinctus USA Louisana 30.500 -91.000 Unknown NORTH-CENT Messenger et al., [12]</td>
</tr>
<tr>
<td>USA USAOPOSSUM cl2 Didelphis marsupialis USA Louisana 30.500 -91.000 Unknown NORTH-CENT Messenger et al., [12]</td>
</tr>
<tr>
<td>9354 9354 Homo sapiens Venezuela Sucre 10.460 -63.610 1999 TcIDOM Messenger et al., [12]</td>
</tr>
<tr>
<td>11541 11541 Homo sapiens Venezuela Merida 8.590 -71.230 2003 TcIDOM Messenger et al., [12]</td>
</tr>
<tr>
<td>11804 11804 Homo sapiens Venezuela Portuguesa 9.084 -69.103 2003 TcIDOM Messenger et al., [12]</td>
</tr>
<tr>
<td>10462P2C3 10462P2C3 Homo sapiens Venezuela Miranda 10.266 -66.485 Unknown TcIDOM This study</td>
</tr>
<tr>
<td>10462P2C7 10462P2C7 Homo sapiens Venezuela Miranda 10.080 -66.449 Unknown TcIDOM This study</td>
</tr>
<tr>
<td>10968P1C1 10968P1C1 Homo sapiens Venezuela Sucre 10.406 -63.298 Unknown TcIDOM This study</td>
</tr>
<tr>
<td>ANT3P1C6 ANTP3C6 Homo sapiens (oral) Venezuela DC 10.500 -66.951 Unknown SOUTH This study</td>
</tr>
<tr>
<td>M16 M16 cl4 Didelphis marsupialis Venezuela Barinas 7.500 -71.230 2004 SOUTH Messenger et al., [12]</td>
</tr>
<tr>
<td>M7 M7 Didelphis marsupialis Venezuela Barinas 7.500 -71.230 2004 SOUTH Messenger et al., [12]</td>
</tr>
<tr>
<td>92122 92122102R Procyon lotor TcIV USA Georgia OUTGROUPS Messenger et al., [12]</td>
</tr>
<tr>
<td>CANIII CANIII cl1 Homo sapiens TcIV Brazil Belen OUTGROUPS Messenger et al., [12]</td>
</tr>
<tr>
<td>CM17 CM17 Dasypus spp. TcIII Colombia Carimagua OUTGROUPS Messenger et al., [12]</td>
</tr>
<tr>
<td>X1060 X1060 cl5 Homo sapiens TcIV Venezuela Guárico OUTGROUPS Messenger et al., [12]</td>
</tr>
<tr>
<td>ERA ERA cl2 Homo sapiens TcIV Venezuela Arrozóategui OUTGROUPS Messenger et al., [12]</td>
</tr>
<tr>
<td>10R26 10R26 Aotus spp. TcIV Bolivia Santa Cruz OUTGROUPS Messenger et al., [12]</td>
</tr>
<tr>
<td>SAIR3 Sairiri3 cl1 Saimiri sciureus TcIV Venezuela Venezuela OUTGROUPS Messenger et al., [12]</td>
</tr>
</tbody>
</table>
sister group relationship with TcI circulating in North America (e.g. [12,13]).

In this manuscript we have set out to evaluate the genetic diversity of TcI in North/Central America, undertaking a comparison with TcI diversity in South America, including TcI_DOM. Our aim was to examine hypotheses describing the origin of the TcI_DOM clade. We propose two possible scenarios: an emergence of TcI_DOM in northern South America as a sister group of North American strains and dispersal among domestic transmission cycles, or an origin in North America, prior to dispersal back into South American domestic cycles, possibly anthropically. To provide further insight into this question we undertook high resolution nuclear and mitochondrial genotyping of multiple Central American strains (from areas of México and Guatemala) and included them in an analysis with other published data [11-13].

A panel of 72 TcI isolates and clones was assembled for analysis (Table 1) [11-16]. Of these, existing sequences and microsatellite data were available for 46 isolates [11,12]. Isolates were classified into three populations: TcI_NORTH-CENT, TcI_SOUTH and TcI_DOM. TcI_NORTH-CENT includes samples from the USA, México, Guatemala and Honduras; TcI_SOUTH corresponds to South America (Argentina, Bolivia, Colombia, Venezuela and Brazil) and TcI_DOM with exclusively domestic isolates from Colombia and Venezuela, already known to correspond to a genotype with restricted genetic diversity: TcLa, as previously described by Herrera et al., (2007) [17] and VENDom, as described by Llewellyn et al., (2009) [13]. Additional DTU isolates (TcIII-TcIV) were included as out-groups in the mitochondrial analysis.

Isolates from México and Guatemala were characterized to DTU level via the amplification and sequencing of glucose-6-phosphate isomerase (GPI) as previously described by Lauthier et al., (2012) [18]. Subsequently, nine maxicircle gene fragments were amplified, sequenced and concatenated from the Mexican and Guatemalan strains according to Messenger et al., 2012 (excluding ND4) [12]. Phylogenetic analysis was also conducted as in Messenger et al., 2012 [12]. Nineteen nuclear microsatellite loci previously described by Llewellyn et al., 2009 [13], were selected based on their level of TcI intra-lineage resolution. Microsatellite loci were amplified across 21 unpublished biological stocks from México and Guatemala. Reaction conditions were as described previously [13]. Dendrograms based on multilocus allele profiles were constructed also according to Llewellyn et al., 2009 [13].

Maxicircle nucleotide diversity (π) was calculated for TcI_NORTH-CENT, TcI_SOUTH and TcI_DOM respectively in DNAsp v5 [19]. Nuclear allelic diversity was calculated for the same populations using allelic richness (A_r) in FSTAT [20]. The resulting values are shown in Figure 1.

Across the 3,449 bp final concatenated alignment (including outgroups), a total of 374 variable sites were found. The mitochondrial phylogeny supported the presence of significant diversity among the isolates examined (Figure 2). TcI_DOM strains formed a monophyletic clade [60% ML BS/0.98 BPP]. The principal division in the
phylogeny was between TcI_SOUTH and TcI_DOM/TcI_NORTH-CENT (98% ML BS/0.98 BPP). However, this division is incomplete, such that a subset of South American strains is also grouped with TcI_DOM and TcI_NORTH-CENT. Thus, it is not possible to conclude that TcI_DOM maxicircle sequences nest uniquely among those from TcI_NORTH-CENT strains. Conversely, a basal relationship of the TcI_NORTH-CENT to TcI_DOM is suggested at the level of nucleotide diversity by population (Figure 1), whereby TcI_DOM < TcI_NORTH-CENT < TcI_SOUTH. Low standard errors about the mean in all three populations, but especially in TcI_DOM and TcI_NORTH-CENT, suggest that sample size had little impact on the accuracy of estimation between populations.

Distance-based clustering using the microsatellite dataset indicated the presence of several well defined clades (Figure 3). Importantly in this case the monophyly of North-Central American isolates was corroborated, and TcI_DOM clustered firmly within them (bootstrap 65%). By contrast, South American isolates fall into a divergent but diverse clade. Thus the nuclear data provide stronger support for divergence of TcI_DOM from within TcI_NORTH-CENT than the maxicircle phylogeny. Sample size-corrected allelic richness estimates are consistent with hierarchical patterns of clustering based on pair-wise genetic distances. As with the maxicircle dataset, there is a pronounced cline in diversity across the populations studied - A_r TcI_DOM < A_r TcI_NORTH-CENT < A_r TcI_SOUTH (Figure 1).
TcI dispersion into Central and North America

Using a 100 MYA biogeographic calibration point [6], molecular clock analyses point to the origin of T. cruzi (sensu stricto) 5 – 1 MYA [21-23] and a most recent common ancestor for TcI at 1.3-0.2 MYA [22]. Reduced genetic diversity among North-Central American isolates by comparison to their southern counterparts is powerful evidence in support of others who suggest that TcI originated in South America [13,24]. The emergence of TcI in the South occurred prior to either migration across the Isthmus of Panama alongside didelphid marsupials during the Great American Interchange [25], or perhaps prior to northerly dispersal via volant mammals (e.g. bats).

Origin of TcIDOM

Recent findings indicate a close resemblance between TcIDOM isolates from the northern region of South America and parasite populations from Central and North America by the use of nuclear and mitochondrial markers [11-13]. Indeed SL-IR genotyping suggests a distribution for TcIDOM that now extends as far south as the Argentine Chaco, where multiple sequences have been identified from human and domestic vector sources [14]. Llewellyn et al., 2009 [13] originally hypothesised that a distinct human/domestic clade could be maintained despite the presence of nearby infective sylvatic strains due to the low parasite transmission efficiency by the vector [13]. In this case multiple feeds by domestic vector nymphs are required to infect individuals, as such human – human transmission is far more common than reservoir host - human transmission. Originally this hypothesis was developed to explain the epidemiology of Chagas disease in Venezuela. However, TcIDOM is clearly widespread and recent data propose a date for its emergence 23,000 ± 12,000 years ago [11].

Figure 3 Isolate grouping of 72 Trypanosoma cruzi I strains based on nineteen nuclear microsatellite markers. Neighbour-joining clustering algorithm implemented. Bootstrap values are included on important nodes. The first figure indicates % bootstrap support over 10,000 trees, the second the % stability over 1000 trees accounting for multi-allelic loci in the dataset. For further details see Llewellyn et al., 2009 [13]. Branch colours indicate isolate origin. The three principal populations TcIDOM, TcISOUTH and TcINORTH-CENT are shown on both map and tree. Red circles correspond to isolates from TcIDOM. Isolates that show clear incongruity between nuclear genotype and maxicircle genotype are marked with reference to Figure 2.
This corresponds to the earliest human colonisation of the Americas [26]. Thus it seems that Tcl_{DOM} may be as ancient as humans in South America. Crucially, our data, which show that Tcl_{DOM} is nested among North Central American strains, suggest that this widespread domestic T. cruzi genotype may actually have made first contact with man in North–Central America.

The expansion of limited diversity genotypes into domestic transmission cycles is a familiar story in T. cruzi. This phenomenon seems to have occurred almost simultaneously with Tcl_{DOM} (<60,000 YA) in the Southern Cone region but involving DTUs TcV and TcVI [22]. Nonetheless, static human population densities sufficient to support a sustained domestic cycle are presumably vital. For Tcl_{DOM} patterns of genetic diversity suggest early colonizing Amerindians may have been responsible for its southerly migration and dispersal from North/ Central America. However, such early settler populations were probably small, dynamic, and inherently unsuitable to sustain transmission of such a genotype. Many questions, therefore, remain unanswered regarding its emergence. Insight could perhaps be drawn from a better understanding of the current distribution and diversity of Tcl_{DOM} (including samples from the Southern Cone), patterns of vector population migrations, and even from analysis of ancient DNA (e.g. [27]). We hope this report serves to galvanize efforts towards this understanding, especially among researchers in Central and North America, where many of the answers lie.

Competing interests

The authors declare no competing financial interests. The funder played no role in the study design.

Authors' contributions

FZE wrote the article, performed the experiments and analysed the data. LAM analysed the data and wrote the article. MAM, TLO, PM, MDL contributed reagents and wrote the paper. CFL analysed the data. JMI, HC, MS contributed reagents. MLS conceived the experiments, analysed the data and wrote the article. MAM, TLO, PM, MDL, MAM, FZE wrote the article, performed the experiments and analysed the data.

Acknowledgements

FZE received an MSc scholarship from the Mexican Council of Science and Technology (CONACyT), and financial support from the International Agency of National Public Health Institutes. LAM, MDL, MAM and MLS acknowledge support from the European FP7 Project ChagasEpiNet, Grant 223034. MLS would like to thank Juan David Ramirez and Prof. Felipe Guhl at the Universidad de los Andes, Colombia for constructive discussion.

Author details

1Centro Regional de Investigación en Salud Pública, Instituto Nacional de Salud Pública, Tapachula, Chiapas, México.
2London School of Hygiene and Tropical Medicine, London, UK.
3Department of Biology, University of Maryland, College Park, MD, USA.
4Area de Entomología Médica, Centro Universitario del Sur, Universidad de Guadalajara, Ciudad Guzmán, Jalisco, México.
5Center for Health Studies, Research Institute, Universidad del Valle de Guatemala, Guatemala City, Guatemala.
6Instituto de Medicina Tropical, Universidad Central de Venezuela, Caracas, Venezuela.

Received: 11 August 2012 Accepted: 3 October 2012 Published: 10 October 2012

References


Submit your next manuscript to BioMed Central and take full advantage of:
- Convenient online submission
- Thorough peer review
- No space constraints or color figure charges
- Immediate publication on acceptance
- Inclusion in PubMed, CAS, Scopus and Google Scholar
- Research which is freely available for redistribution

Submit your manuscript at www.biomedcentral.com/submit

BioMed Central
4.4 Discussion
The three papers presented in this chapter describe the natural population structures of *T. cruzi* TcI in three different epidemiological situations.

In both Bolivia and Brazil, where sylvatic transmission cycles were sampled, considerable genetic diversity was observed among TcI populations, which was consistent with intense, local parasite transmission and/or minimal extinction of distinct genotypes. In many of the study areas, multilocus genotypes (MLGs) were rarely repeated indicating only a fraction of total genetic variation was sampled. Indeed, all three of these studies would have benefited from additional strains. In Brazil, a novel, highly divergent, homogenous clade isolated from multiple different host and vector species was identified, confirming our sampling of intra-TcI genetic diversity has not yet reached saturation (Lima et al., 2014). In Bolivia, a cohort of patient samples from the three highland study sites is needed to corroborate the link between highland domestic and sylvatic gene flow; a putative, hybridization zone in northern Cochabamba also warrants further attention (Messenger et al., accepted). Characterization of new patient isolates from Central America will be critical to define the geographical range of TcI_DOM which currently extends from Venezuela to northern Argentina (Zumaya-Estrada et al., 2012).

In addition to the inherent problems of *T. cruzi* isolation (section 3.1.2), sylvatic sampling is also logistically difficult and intrinsically biased. Certain species are notoriously overrepresented in trapping efforts, e.g. *Didelphis morphina*, and due to their characteristically high circulating parasitaemias, facilitate greater parasite isolation rates (Legey et al., 2003; Llewellyn et al., 2009a). The importance of a particular host to disease ecology will depend on the balance between *T. cruzi* prevalence and relative local abundance; if a species with high infection rates represents only a negligible proportion of the total population then its role in transmission is lesser compared to a second species that displays a lower disease prevalence but is significantly more abundant in the area.

An improved understanding of sylvatic transmission dynamics would be advantageous both in terms of Chagas disease control and as well as habitat conservation. Serological surveillance of targeted peridomestic populations has been initiated in Brazil (Jansen and Roque, 2010), Argentina (Gürtler et al., 2007), Venezuela (Crisante et al., 2006), Mexico (Estrada-Franco et al., 2006) and the USA (Tenny et al., 2014), using these hosts as bioindicators (sentinels) of *T. cruzi* transmission risk, considering high seroprevalence in these species usually precedes the emergence of local human Chagas disease (Roque et al., 2008; Jansen and Roque, 2010). In parallel, detectable reductions in parasite genetic diversity as a direct result of human disruption, highlights the potential to exploit such measurements as proxies for overall ecosystem health (Messenger et al., accepted; Lima et al., 2014).

The three papers included in this chapter demonstrate the significant impact human activity has had on the evolution of *T. cruzi*. Direct habitat destruction, especially deforestation, provides ample opportunities for synanthropic mammalian species (especially marsupials, rodents and bats) and triatomines, to colonize peridomestic areas seeking shelter and food (Walsh et al., 1993; Abad-Franch and Monteiro, 2007). The corresponding reduction in
parasite genetic diversity, has potential implications for human Chagas disease with respect to parasite virulence, transmissibility and drug susceptibility. Future studies could incorporate in vitro and in vivo characterization of these homogeneous strains in comparison with their ‘undisturbed’ sylvatic counterparts, to determine whether these isolates possess a phenotypic advantage. It seems likely that such a phenomenon may exist with TcI.Dom strains, considering their historical maintenance among small, highly dynamic human populations, and extensive contemporary geographical distribution.

Unlike genetic exchange in other typanosomatid species, which involves Mendelian segregation of parental alleles (Jenni et al., 1986; Peacock et al., 2011; Akopyants et al., 2009), T. cruzi is known to possess an unorthodox set of potentially alternate recombination mechanisms (Gaunt et al., 2003; Lewis et al., 2011; Messenger et al., 2012; Ramirez et al., 2012; Baptista et al., 2014). Mitochondrial introgression was detected among a minority of strains in all three studies and appears to be a natural, non-obligatory occurrence among TcI populations (Messenger et al., 2012; Ramirez et al., 2012).

In Bolivia, introgression was observed between arboreal populations, in Brazil within a single isolate from D. albiventris in Caatinga, and in the TcI.Dom study, one human strain in Mexico. No reciprocal nuclear hybridization was identified among these recombinant isolates, which may reflect a mating system involving asymmetric mitochondrial inheritance. Considering the fundamental role T. cruzi maxicircles play in parasite metabolism and development in the triatomine vector and the observation that these parasite populations were principally structured by ecological fitting, it is not implausible to suggest that such a mechanism might facilitate host range extension and/or resource tracking. Similarly, in Bolivia, excess heterozygosity was observed among highland strains, potentially indicative of another recent hybridization event (Messenger et al., accepted). Alternatively, sustained, conserved heterozygosity has been proposed to provide an adaptive fitness advantage in triatomine bugs through metabolic flexibility over a range of environmental temperatures (Widmer et al., 1987). Both introgression and heterozygous recombinant strains are also candidates for in vitro and in vivo characterization to elucidate the impact of hybridization on T. cruzi phenotype.

The occurrence of non-canonical, alternate mating systems in T. cruzi complicates the interpretation of many conventional population genetics indices (Ramírez and Llewellyn, 2014). The use of model-based population assignment software, e.g. STUCTURE (Pritchard et al., 2000) and BAPS (Corander et al., 2003), is contraindicated as these programs are based on algorithms which assume Hardy-Weinberg allele frequencies and complete linkage equilibrium between genetic markers, two criteria that are largely violated by clonal reproduction in T. cruzi. With regards to detecting hybridization, sample population allocation is crucial; grouping of divergent non-recombining subgroups (in the case of T. cruzi, distinct DTUs) can inflate genetic linkage statistics and mask recombination events occurring between more closely related individuals (Smith et al., 1993). Recent observations of the Wahlund effect obscuring Hardy-Weinberg allele frequencies and linkage equilibrium within Brazilian TcII strains, caution the interpretation of statistics derived from inappropriately assigned parasite populations (Baptista et al., 2014).
To circumvent these limitations, in the Bolivia study, natural parasite groupings were defined by submitting clones a priori to two different clustering methodologies, both independent of Hardy-Weinberg assumptions, and congruence was compared between outputs. This strategy is not ideal and highlights the need to develop new statistical tools and/or adapt those used to define population structures in other organisms with more covert unorthodox mating systems (Hickman et al., 2013; Ni et al., 2013).

Few studies, including those described herein, have had the capacity and expertise to undertake intensive, representative parasite isolation from all available mammalian and vector species in a local ecosystem. An integrated landscape genetics approach (Biek and Real, 2010; Manel and Holderegger, 2013), measuring geospatial biotic (e.g. local fauna diversity, relative species abundances, etc.) and abiotic parameters (e.g. temperature, humidity, water availability etc.), complemented by improved population genetics analyses, may represent a promising new strategy to examine the complex interplay between T. cruzi, its mammalian hosts and ecological niches.
5. Detection of genetic exchange among natural *T. cruzi* populations

5.1 Background

5.1.1. Natural genetic exchange in *T. cruzi*

The principal mode of reproduction among a number of parasitic protozoan species is the subject of an ongoing, intense, decades-old debate (Tibayrenc *et al*., 1990; Tibayrenc and Ayala, 1991; 2012; 2013; 2014a; 2014b; Ramírez and Llewellyn, 2014; Tomasini *et al*., 2014a; 2014b). At the two extremes are the preponderate clonal evolution (PCE) model, which suggests that genetic exchange is too infrequent to break the predominant pattern of clonality, such that only ‘restrained recombination’ occurs on an evolutionary scale (Tibayrenc and Ayala, 2012; 2013; 2014a; 2014b), and the counter-proposition that hybridization is pervasive, albeit challenging to detect, among some natural disease foci (Ramírez and Llewellyn, 2014). *T. cruzi* often fulfills key assumptions of PCE, namely strong linkage disequilibrium (LD), deviations from Hardy-Weinberg allele frequencies and structuring of populations into stable, discrete genetic clusters, all of which have been cited as compelling evidence that it is essentially a clonal organism (Tibayrenc and Ayala, 2012; 2013).

With improved sampling strategies and the use of more resolutive genotyping techniques, a growing amount of field data now indicate that natural recombination in *T. cruzi* may be frequent, non-obligatory and idiosyncratic, potentially involving independent exchange of kinetoplast and nuclear genetic material, as well as canonical meiotic mechanisms (Table 5.1). At the inter-lineage level, DTUs TcV and TcVI are unequivocal hybrids of TcII and TcIII, which resemble diploid Mendelian F1 progeny, sharing intact alleles from their parental strains (Machado and Ayala, 2001; Brisse *et al*., 2003; Barnabé *et al*., 2011; Lewis *et al*., 2009b; 2011; Yeo *et al*., 2011). The status of TcIII and TcIV as ancient recombinants is more contentious, but supported by some nuclear markers (Westenberger *et al*., 2005) and the sharing of mitochondrial haplotypes (Lewis *et al*., 2011; Messenger *et al*., 2012).

At the intra-lineage level, genetic exchange is increasingly reported, particularly among TcI populations, but it is unclear whether this is due to the examination of isolates that are minimally subdivided spatially and temporally, and therefore more likely to undergo hybridization, or whether it truly reflects the analysis of strains that are more permissive to recombination (Ramírez and Llewellyn, 2014). The underlying cytological mechanisms of natural intra-TcI recombination are unresolved and vary between studies and genetic markers used (Table 5.1).

Mitochondrial introgression is emerging as a common feature of natural transmission cycles especially within TcI populations (Messenger *et al*., 2012; accepted; Zumaya-Estrada *et al*., 2012; Lima *et al*., 2014) but also between major lineages (Lewis *et al*., 2011; Messenger *et al*., 2012; Barnabé and Breniere, 2012; Roellig *et al*., 2013) (Table 5.1). The evidence for
unequivocal reciprocal nuclear recombination among introgression strains is thus far lacking. One explanation, given their role in growth, development and metabolism, is that asymmetric mitochondrial introgression may facilitate host range extension, satisfying the elevated necessity to escape Muller’s ratchet compared to the nuclear genome (Messenger et al., accepted; Neiman and Taylor, 2009; Ramírez and Llewellyn, 2014). However, these observations have been interpreted by others as attributable to gross differences in evolutionary pressures and molecular clocks between non-coding microsatellites and coding maxicircle genes (Tibayrenc and Ayala, 2013).

It is clear that to improve detection of natural recombination, the use of multiple, different types of molecular markers (nuclear and mitochondrial, coding and non-coding) are required in combination with targeted investigation of potential ‘hybridization’ zones, i.e. areas where recently diverged, genetically distinguishable subpopulations come into regular contact (Messenger et al., accepted; Ramírez and Llewellyn, 2014). The value of such high density sampling has already been demonstrated in defining the population structures of other trypanosomatid species, e.g. *T. b. gambiense* (Koffi et al., 2009), *T. congolense* (Morrison et al., 2009a), *L. braziliensis* (Rougeron et al., 2009) and *L. guyanensis* (Rougeron et al., 2011), including establishing putative levels of genetic exchange.
<table>
<thead>
<tr>
<th>Reference</th>
<th>Population(s)</th>
<th>Location</th>
<th>Type of Genetic Markers</th>
<th>Evidence of Genetic Exchange</th>
<th>Putative Mechanism</th>
<th>Supporting Evidence</th>
</tr>
</thead>
<tbody>
<tr>
<td>Messenger et al., 2012</td>
<td>TcI</td>
<td>Domestic</td>
<td>MLMT</td>
<td>Intra-lineage</td>
<td>MLEE</td>
<td>Putative homozygous parents and Hardy-Weinberg phenotypes</td>
</tr>
<tr>
<td>Barnabé and Breniere, 2012</td>
<td>TcI</td>
<td>Domestic</td>
<td>MLMT</td>
<td>Intra-lineage</td>
<td>MLEE</td>
<td>Recombinant mitochondrial sequence with no detectable nuclear involvement</td>
</tr>
<tr>
<td>Caruso et al., 1996</td>
<td>TcI</td>
<td>Domestic</td>
<td>RAPD, GPI</td>
<td>Intra-lineage</td>
<td>MLEE</td>
<td>Recombinant mitochondrial introgression between TcI &amp; TcII/III/IV/V with no detectable nuclear involvement</td>
</tr>
<tr>
<td>Ocaño et al., 2010</td>
<td>TcI</td>
<td>Domestic, sylvatic</td>
<td>MLMT</td>
<td>Inter-lineage</td>
<td>MLEE</td>
<td>Mitochondrial introgression between TcI &amp; TcIII/IV/V/VI with no detectable nuclear involvement</td>
</tr>
<tr>
<td>Carrasco et al., 2012</td>
<td>TcI</td>
<td>Domestic</td>
<td>RAPD, GPI</td>
<td>Intra-lineage</td>
<td>MLEE</td>
<td>Mitochondrial introgression between TcI &amp; TcIII/IV/V/VI with no detectable nuclear involvement</td>
</tr>
</tbody>
</table>

**Table 5.1. Summary of field evidence of genetic exchange in T. cruzi.**
<table>
<thead>
<tr>
<th>Year</th>
<th>Location</th>
<th>Inversion Type</th>
<th>Loci</th>
<th>Mitochondrial Introgression</th>
<th>Nuclear Involvement</th>
<th>Other Notes</th>
</tr>
</thead>
<tbody>
<tr>
<td>2013</td>
<td>TcI &amp; TcIII/IV</td>
<td>Inter-lineage</td>
<td>COI, ND1</td>
<td>Mitochondrial introgression between TcI &amp; TcIII/IV</td>
<td>No detectable nuclear</td>
<td>Asymmetrical mitochondrial introgression</td>
</tr>
<tr>
<td>2012</td>
<td>TcI &amp; TcII</td>
<td>Inter-lineage</td>
<td>ND4, ND7</td>
<td>Mitochondrial introgression between TcI &amp; TcII</td>
<td>No detectable nuclear</td>
<td>Asymmetrical mitochondrial introgression</td>
</tr>
<tr>
<td>2011</td>
<td>Lewis et al.</td>
<td>Inter-lineage</td>
<td>COI</td>
<td>Mitochondrial introgression between TcI &amp; TcII</td>
<td>No detectable nuclear</td>
<td>Asymmetrical mitochondrial introgression</td>
</tr>
<tr>
<td>2014</td>
<td>Baptista et al.</td>
<td>Intra-lineage</td>
<td>24S α rRNA, 18S rRNA, TcMSH2, Tc55, DHFR, TS, COI, ND1</td>
<td>Mitochondrial introgression between TcI &amp; TcIII/IV</td>
<td>Intra-lineage</td>
<td>Asymmetrical mitochondrial introgression</td>
</tr>
<tr>
<td>2013</td>
<td>Bankole et al.</td>
<td>Intra-lineage</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>
5.1.2. *In vitro recombination in T. cruzi*

The generation of intra-TcI hybrids *in vitro* strongly support the premise that at least some *T. cruzi* strains have an extant capacity for genetic exchange (Gaunt *et al.*, 2003). The putative parental isolates identified by Carrasco *et al.*, were transformed with episomal recombinant plasmids containing either hygromycin B or neomycin resistance genes and co-passaged through *in vitro* (mammalian cell cultures) and *in vivo* (mice and triatomine bugs) life cycles (Gaunt *et al.*, 2003). Isolation of six clones by double drug selection from *in vitro* axenic cultures, and subsequent genetic characterization by MLEE, karyotyping, microsatellites and nucleotide sequencing of housekeeping genes, demonstrated that these intra-lineage recombinants had inherited all parental alleles at most loci and one parental maxicircle genotype.

By analogy with *Candida albicans* (Bennett and Johnson, 2003; Forche *et al.*, 2008), it was proposed that nuclear fusion had created a tetraploid intermediate, followed by homologous recombination, gradual genome erosion and reversion to aneuploidy. FACs analysis of hybrid isolates indicated a stable DNA content, on average, 69% higher than parental strains (Lewis *et al.*, 2009b). Subsequent prolonged maintenance in axenic cultures demonstrated a gradual, progressive decline in DNA content, with no evidence of any true meiotic reductive division; to date these experimental hybrids remain sub-tetraploid (Lewis *et al.*, 2010).

While this parasexual mechanism of genetic exchange has a precedent in fungal species, it is challenging to reconcile with both the patterns of allele inheritance observed among natural *T. cruzi* populations (Table 5.1) as well as the conservation of meiosis-specific orthologues within the *T. cruzi* genome (Ramesh *et al.*, 2005). A similar paradox exists in *T. b. brucei* where canonical meiotic recombination (Peacock *et al.*, 2011), including the formation of haploid life cycle stages (Peacock *et al.*, 2014), has been explicitly described *in vitro*, but is not the exclusive mechanism reported from transmission cycles (Duffy *et al.*, 2013). Likewise, experimental hybridization in *Leishmania* resembles meiosis (Akopyants *et al.*, 2009; Inbar *et al.*, 2013) but both aneuploidy and inbreeding are frequent in nature (Rougeron *et al.*, 2009; 2011; Sterkers *et al.*, 2011; 2014; Calvo-Álvarez *et al.*, 2014; Rogers *et al.*, 2014).
5.2 Objectives
The aim of this chapter was to exploit the phylogenetic markers developed in chapter 3 to measure the frequency of natural genetic exchange, identify any potential underlying mechanisms and examine its impact on *T. cruzi* population structuring.

Specific objectives were to:

a. Investigate the principal mating strategy of TcI within an intensively-sampled endemic disease focus in Colombia.

b. Compare incongruence between nuclear and mitochondrial topologies to uncover mitochondrial introgression events occurring among TcI populations.

c. Examine the inheritance patterns of natural hybrids strains from Colombia to resolve their putative status as novel recombinants.

d. Characterize intra-TcV and TcVI genetic diversity across its expanding geographical range.
5.3 Results

5.3.1 Cryptic sexuality in T. cruzi

The majority of field evidence suggests T. cruzi does not conform to strict clonality (Tibayrenc and Ayala, 1991) or panmixia (Harvey and Keymer, 1987) and that recombination is frequent, non-obligate and idiosyncratic, potentially involving independent exchange of kinetoplastid and nuclear genetic material (Carrasco et al., 1996; Machado and Ayala, 2001; Ocaña-Mayorga et al., 2010; Messenger et al., 2012; Lewis et al., 2011; Roellig et al., 2013; Baptista et al., 2014). However, the relative contributions of alternate mating strategies to T. cruzi population structures are still strongly debated (Lewis et al., 2011; Tibayrenc and Ayala, 2012; 2013; Ramirez and Llewellyn, 2014).

To examine the predominant mode of parasite reproduction among natural populations, high resolution nuclear (Llewellyn et al., 2009a) and mitochondrial (Messenger et al., 2012) genotyping was used to characterize 269 TcI biological clones isolated from domestic, peridomestic and sylvatic transmission cycles in three Colombian provinces (Boyaca, Casanare and Santander).

This study is reported in full below in Ramírez, et al. 2012.

In summary:

- Measurements of $A_r$ demonstrated uniformly high genetic diversity across all study sites ($A_r=2.958\text{-}3.778$) with the exception of a sylvatic population in Boyaca ($A_r=1.633$), which was likely attributable to local human-mediated habitat degradation.
- MLGs were rarely repeated which was consistent with intense local parasite transmission. Strongly significant multilocus linkage disequilibrium (MLD) ($I_A = p<0.0001$ for all populations) indicated that clonal propagation was the predominant mode of nuclear reproduction.
- Non-parametric nuclear clustering and phylogenetic analyses supported a clear genetic subdivision between domestic and sylvatic populations; most peridomestic strains were poorly differentiated from those found in local sylvatic environments. The majority of domestic isolates were robustly grouped together (TcIDOM).
- A subset of Colombian patients were ‘super-infected’, i.e. harbored a mixture of different clones, some of which had sylvatic origins. Isolation of only TcI from a neonate born to a co-infected TcI-TcII mother, suggests parasite DTU may also influence congenital T. cruzi transmission.
- Gross nuclear-mitochondrial phylogenetic incongruence identified multiple, independent mitochondrial introgression events among ~20% of isolates, challenging clonality as the predominant mating strategy in Colombia. No evidence of reciprocal nuclear recombination was observed among any recombinant strains. This may reflect an alternate mating system involving asymmetric mitochondrial inheritance or the relative sampling amount of each genome (20% of the mitochondrial genome vs. <0.1% of the nuclear genome).
- A mosaic maxicircle sequence was detected in a human isolate from Santander by the presence of a recombination breakpoint within a contiguous region of the MURF1 gene, and confirmed by allele-specific PCR, to exclude PCR error and/or Taq polymerase template switching. Inter-molecular mitochondrial recombination represents another method of generating novel genetic diversity and may function to prevent the accumulation of deleterious mutations following clonal expansion (‘Muller’s Ratchet’).

- Bayesian skyline plots (BSPs) constructed from mitochondrial sequence data indicated that Colombian TcI_{DOM} strains emerged 23,000 ± 12,000 years ago, and underwent population expansion, broadly coinciding with the earliest human migration into South America.
COVER SHEET FOR EACH ‘RESEARCH PAPER’ INCLUDED IN A RESEARCH THESIS

Please be aware that one cover sheet must be completed for each ‘Research Paper’ included in a thesis.

1. For a ‘research paper’ already published

1.1. Where was the work published? Molecular Ecology  

1.2. When was the work published? 2012  

1.2.1. If the work was published prior to registration for your research degree, give a brief rationale for its inclusion  

N/A  

1.3. Was the work subject to academic peer review? Yes  

1.4. Have you retained the copyright for the work? Yes / No  

- If yes, please attach evidence of retention.  
- If no, or if the work is being included in its published format, please attach evidence of permission from copyright holder (publisher or other author) to include work

2. For a ‘research paper’ prepared for publication but not yet published

2.1. Where is the work intended to be published?  

2.2. Please list the paper’s authors in the intended authorship order  

2.3. Stage of publication – Not yet submitted / Submitted / Undergoing revision from peer reviewers’ comments / In press

3. For multi-authored work, give full details of your role in the research included in the paper and in the preparation of the paper. (Attach a further sheet if necessary)

The candidate had significant input into experimental design, contributed laboratory reagents and reference datasets, supervised the generation of mitochondrial data, and participated in data analysis and drafting of the final manuscript.

NAME IN FULL (Block Capitals) LOUISA ALEXANDRA MESSENGER  

STUDENT ID NO: 223021  

CANDIDATE’S SIGNATURE  

Date 10/2/15  

SUPERVISOR/SENIOR AUTHOR’S SIGNATURE (3 above)
Contemporary cryptic sexuality in *Trypanosoma cruzi*

**Juan David Ramírez,** *Felipe Guhl,* **Louisa A. Messenger,** *Michael D. Lewis,* **Marleny Montilla,** *Zulma Cucunuba,* **Michael A. Miles** and **Martin S. Llewellyn**

*Centro de Investigaciones en Microbiología y Parasitología Tropical – CIMPAT, Universidad de los Andes, Bogotá, Colombia,*

*Grupo de Parasitología, Instituto Nacional de Salud – INS, Bogotá, Colombia,*

*Department of Pathogen Molecular Biology, London School of Hygiene and Tropical Medicine, London, UK*

**Abstract**

Clonal propagation is considered to be the predominant mode of reproduction among many parasitic protozoa. However, this assumption may overlook unorthodox, infrequent or cryptic sexuality. *Trypanosoma cruzi*, which causes Chagas disease, is known to undergo non-Mendelian genetic exchange in the laboratory. In the field, evidence of extant genetic exchange is limited. In this study, we undertook intensive sampling of *T. cruzi* Discrete Typing Unit I in endemic eastern Colombia. Using Fluorescence-activated cell sorting, we generated 269 biological clones from 67 strains. Each clone was genotyped across 24 microsatellite loci. Subsequently, 100 representative clones were typed using 10 mitochondrial sequence targets (3.76 Kbp total). Clonal diversity among humans, reservoir hosts and vectors suggested complex patterns of superinfection and/or coinfection in oral and vector-borne Chagas disease cases. Clonal diversity between mother and foetus in a congenital case demonstrates that domestic TcI genotypes are infective in utero. Importantly, gross incongruence between nuclear and mitochondrial markers is strong evidence for widespread genetic exchange throughout the data set. Furthermore, a confirmed mosaic maxicircle sequence suggests intermolecular recombination between individuals as a further mechanism of genetic reassortment. Finally, robust dating based on mitochondrial DNA indicates that the emergence of a widespread domestic TcI clade that we now name TcI<sub>DOM</sub> (formerly TcI<sub>a</sub>/VEN<sub>DOM</sub>) occurred 23 000 ± 12 000 years ago and was followed by population expansion, broadly corresponding with the earliest human migration into the Americas.

**Keywords:** disease biology, empirical evolution of sex, molecular evolution, parasitology, population genetics, protists

Received 20 March 2012; revision received 16 May 2012; accepted 25 May 2012

**Introduction**

Several models exist to describe the population genetic structure of parasitic organisms. At the extremes of the spectrum are panmixia (Harvey & Keymer 1987) and pure clonality (Tibayrenc & Ayala 1991), but also several intermediates exist, tailored to particular epidemiological scenarios within the same species (Maynard Smith *et al.* 1993). Population data collected from the field, as well as experimental crosses in the laboratory, now suggest that such intermediates provide the ‘best-fit’ for many parasitic protozoa. Occasional (or pervasive) sex and associated random segregation between genetic markers are reported variously among *Plasmodium, Giardia, Leishmania, Toxoplasma, Trypanosoma brucei* ssp. and *Trypanosoma cruzi* (Jenni *et al.* 1986; Grigg & Suzuki 2003; Mzilahowa *et al.* 2007; Akopyants *et al.* 2009; Birky 2011; Ocana-Mayorga *et al.* 2011). However, while laboratory crosses demonstrate a capacity for genetic exchange, field evidence from allele frequencies among natural populations can be limited and equivocal. Parasite population geneticists commonly rely on rapidly evolving short tandem repeat (STR) loci at local
spatial and temporal scales to detect genetic exchange (Rougeron et al. 2009; Ocana-Mayorga et al. 2011). However, data interpretation is hampered by blunt statistical tools whose models deal poorly with partial clonality and/or non-Mendelian genetic exchange at a population level. Important theoretical advances have been made in the interpretation of such statistics (De Meus et al. 2006). However, complementary approaches examining extranuclear genomes are required to bolster such data and provide compelling evidence that sex among some species is extant, frequent and epidemiologically important, even if it does not conform to classic Hardy–Weinberg expectations.

Trypanosoma cruzi is a zoonotic kinetoplastid protozoan and the agent of Chagas disease in Latin America. Transmission to humans occurs mostly via the infected faeces of domiciliated triatomine species. However, important nonvectoral transmission routes are also recognized, including oral transmission via contaminated fruit products and congenital transmission.

Structured genetic diversity in other medically important trypanosomatids has led to the designation of species [Leishmania sp. (Miles et al. 2009)] or subspecies [T. brucei spp. (Hide & Tait 2009)]. Similarly, divergent T. cruzi taxonomic units are defined as discrete typing units (DTUs). Six DTUs have been described: Tcl–TcVI (Zingales et al. 2009). The distribution of these DTUs is related to several ecological and epidemiological variables, including host, geography and transmission cycle [wild (henceforth ‘sylvatic’) vs. domestic]. Among the DTUs, Tcl is the most abundant and widely dispersed. Tcl has also been the focus of recent attempts to characterize genetic diversity at the sub-DTU level and several epidemiologically important findings have emerged (Herrera et al. 2007; Llewellyn et al. 2009; Ocana-Mayorga et al. 2011; Ramírez et al. 2011a). The ability to isolate multiple T. cruzi clones from individual hosts means that patterns of superinfection and infrapopulation diversity within Tcl can now be examined (Llewellyn et al. 2011).

A defining feature of kinetoplastid protozoa is the kinetoplast, the organelle which contains the trypanosome’s mitochondrial (mt) DNA. Kinetoplast DNA comprises a network of tens of thousands of interlocking c.1 kb minicircles, interlaced with 20–50 larger (c. 25 Kb) maxicircles. In T. cruzi, minicircle populations are highly heterogenous within a single clone (Tellera et al. 2006). However, next generation sequencing reveals that maxicircles are relatively homogeneous by comparison, at least within their coding region (Messenger et al. 2012). This feature, and the availability of three complete maxicircle sequences, has enabled mtDNA multilocus sequence typing of T. cruzi and revealed minor coding heteroplasmy (Messenger et al. 2012). Among animals, the mtDNA mutation rate is an order of magnitude higher than that of the nuclear genome (Haag-Liautard et al. 2008). A similar mutation rate differential is apparently true for T. cruzi (Machado & Ayala 2001; Messenger et al. 2012), with corresponding power to distinguish closely related clones.

On the basis of limited laboratory (Gaunt et al. 2003) and field (Machado & Ayala 2001) data, maxicircle inheritance in T. cruzi is thought to be uniparental. Several studies have now shown evidence for cases of interlineage mitochondrial introgression, whereby the maxicircle genome from one DTU is present alongside an apparently nonrecombinant nuclear genome from a different DTU (Machado & Ayala 2001; Messenger et al. 2012). Clear incongruence between nuclear and mitochondrial phylogenies results.

In this study, we evaluated diversity at 24 nuclear STR loci across 269 Tcl clones and 10 mitochondrial sequence loci (3.76 Kb) across a subset of 100 clones, most from a restricted disease focus in Colombia. We identified frequent incongruence between nuclear and mitochondrial clustering as the first robust evidence that Tcl is undergoing genetic exchange repeatedly in natural transmission cycles. Furthermore, Bayesian skyline plots (BSPs) of an expanded mtDNA data set indicated that the emergence of a widespread domestic clade (TcI DOM) occurred c. 23 000 years ago, broadly corresponding with the first human arrivals in South America.

Methods

Study area, parasite cloning and Trypanosoma cruzi genotyping

Three provinces (Boyaca, Casanare and Santander) were sampled in north-eastern Colombia (Fig. S1, Supporting information). Multiple triatomines and mammals were captured at domestic (within dwellings), peridomestic (near dwellings) and sylvatic (>10 m from dwellings) sites. Trypanosomes from patients were isolated after written informed consent and with local ethical clearance approved by National Institute of Health (CTIN 012/08). For a full list of isolates and clones see Table S1 (Supporting information). Parasites were cloned from culture either into Liver Infused Trypomastigote containing wells on 96-well plates using a BD FACSAria flow cytometer or via limiting dilution (Ramírez et al. 2011b). Clones for analysis were selected at random among those that grew in the plates. DNA was extracted using a Qiamp DNA isolation kit. Trypanosoma cruzi DTU was preliminarily defined via a 400 bp glucose phosphate isomerase (GPI) gene fragment (primers in Table S3, Supporting information), for which a wealth of
comparative data are available from the literature (Brou- 
tin et al. 2006; Lewis et al. 2011). Two hundred and sev-
enty sequences were deposited in GenBank.

Microsatellite analysis
Twenty-four microsatellite loci, distributed across eight 
putative chromosomes, were amplified as described 
previously (Llewellyn et al. 2009) (Table S2, Supporting 
information). Prior to population genetic analysis, iso-
lates were grouped a priori into seven populations: Boy-
aca domestic (BD), Boyaca peridomestic (BP), Boyaca 
ylvatic (BS), Casanare domestic (CD), Casanare perido-
mestic (CP), Casanare sylvatic (CS) and Santander 
omic (SD). Statistics were calculated for each popu-
lation as previously (Table S3, Supporting information) 
(Ocana-Mayorga et al. 2011).

Individual level sample clustering was defined via a 
neighbour-joining tree based on pairwise distances 
between multilocus genotypes MLGs [evaluated using 
$D_{AS}$ (1 – proportion of shared alleles at all loci/n)] cal-
culated in MICROSAT (Minch et al. 1995). To define a 
posteriori the number of putative populations in the data 
set using a nonparametric (free from Hardy-Weinberg 
constraints) approach, we employed a K-means cluster-
ing algorithm, implemented in adegenet. As described 
in Jombart et al. 2010; the ’true’ number of populations 
can be defined by reference to the Bayesian Information 
Criterion (BIC), which reaches a minimum when the 
best supported assignment of individuals to the appro-
priate number of clusters is approached Jombart et al. 
2010. In practice, this number is selected at the ’elbow’ 
of the BIC curve (Fig. 1). The relationship between 
these clusters and the individuals within them can be 
evaluated via a discriminant analysis of principal com-
ponents (DAPC), again as in (Jombart et al. 2010). We 
chose to retain the number of principal components 
(PCs) that represented the first 80% of the total varia-
tion in the data set. DAPC results are presented as mul-
tidimensional scaling plots in Figs 1 and 2.

Multilocus mtDNA (maxicircle) sequence analysis
One hundred clones representative of the total nuclear 
diversity were selected for multilocus mtDNA sequence 
analysis. Ten mitochondrial maxicircle DNA fragments 
were sequenced and concatenated as in Messenger et al. 
(2012) (Table S2, Supporting information). This data set 
was evaluated in MODELTEST 3.7 (Posada & Crandall 
1998) where the most appropriate evolutionary model 
was selected based on the Akaike Information Criterion. 
A maximum composite likelihood analysis using a 
Tamura-3 parameter model and the neighbour-joining 
algorithm was run in RAxML 7.2.5. To evaluate the 
robustness of the nodes in the resulting phylogenetic 
tree, 1000 bootstrap replicates were performed. The 
final tree was rooted with Esmeraldo (TcII) and CANIII 
(TcIV) sequences. Sequences were deposited in Gen-
Bank.

The mtDNA data set was submitted to RDP (Martin 
et al. 2005) to identify potential mosaic recombinants. 
Several algorithms were implemented including GEN-
ECONV, BOOTSCAN/RECSCAN and MaxChi. Putative 
mosaics were confirmed by eye and empirically via alle-
lic specific PCR (ASP) using specially designed primers 
for regions flanking the putative breakpoints (Fig. S3, 
Supporting information).

Molecular dating using mtDNA sequences
To provide a date for the emergence of the key human 
TcI clade, we used a Bayesian Markov chain Monte 
 Carlo (MCMC) method implemented in the Bayesian 
evolutionary analysis by sampling trees (BEAST) pack-
age (Drummond et al. 2005). Excluding three divergent 
and/or recombinant clones, a final 3.7 Kbp alignment of 
concatenated maxicircle sequence from 97 TcI single-
celled clones was assembled for analysis. We included 
nine further TcI/TcIV isolates from Messenger et al. 
(2012) including several from North/Central America, 
as well as from the TcIDOM clade (VEN/Dom/Tca) (Lle-
wellyn et al. 2009; Herrera et al. 2007). A full list of 
samples used is included in Table S1 (Supporting infor-
mation). Further details of methodology are included in 
Table S5 (Supporting information).

Results
Genetic diversity, heterozygosity and linkage
Two hundred and sixty-nine clones were genotyped 
against 24 microsatellite loci (Table S4, Supporting 
information) and we identified 212 distinct MLGs. Esti-
mates of allelic richness indicate comparable and high 
genetic diversity among populations (BD, BP, CD, CP, 
CS, SD, $A_r = 2.958–3.778$, Table S3, Supporting infor-
mation), but showed a marked reduction in BS (1.633, 
Table S3, Supporting information). Over loci, $F_{IS}$ values 
indicate a widespread deficit in heterozygosity, corrobo-
rated by significant deviation from H-W allele frequen-
cies at individual loci ($F_{IS} = 0.37–0.17$, 55–52% 
polymorphic loci with significant deficit in heterozygos-
ity, Table S3, Supporting information). Again, BS repres-
sents a clear outlier with strongly negative values for 
$F_{IS}$ (~0.65, Table S3, Supporting information) and corre-
sponding high heterozygosity (41% of loci with signifi-
cant excess). Multilocus linkage disequilibrium, manifest in strongly significant values for the $I_A$ across
all sufficiently diverse populations, suggests (misleadingly) that clonal reproduction prevails throughout the data set.

Across the final, 3680 bp concatenated maxicircle fragment, we identified 243 variable sites within TcI.

Clustering based on nuclear data

Thirty clusters were defined among the 269 clones subjected to DAPC, once 41 PCs had been retained (80% of the total variation) and subjected to analysis. A full list of the assignments of individual clones to DAPC populations is included in Table S1 (Supporting information). A neighbour-joining (NJ) tree was constructed from the same data set for comparison (Fig. 1). Broad congruence was observed between the two methodologies; however, DAPC provided the additional benefit of allowing better visualization of the relative distances between groups. As such, the majority of populations made up of nonhuman isolates tended to group together (most notably those defined by the label ‘x’, Fig. 1). However, among sylvatic clones some clear outliers are evident, most notably clones from BS (pale green, NJ Tree, Fig. 1) and a subset (group 20) from CS.

Fig. 1 Genetic clustering among 269 Colombian TcI clones including oral, congenital and vector borne human disease cases. Top – Multidimensional scaling plot is based on a DAPC for 30 clusters defined via K-means clustering (10⁹ iterations, 41 PCs (80% of variation) retained, BIC curve inserted – error bars represent SD about the mean of five independent runs). Clusters are defined by distinct colours and inertia ellipses, dots represent individual clones. Clusters 14, 15 and 16 contain clones defined as belonging to nuclear Cluster 6 (nuclear) in Fig. 2. Bottom – Unrooted neighbour-joining tree constructed using individual pairwise $D_{AS}$ values. Closed circles indicate clades with bootstrap stability ≥70% across 1000 replicates. Branch colours correspond to province and transmission cycle: red – Casanare domestic, pink – Boyaca domestic, orange – Santander domestic (SD), dark blue – Casanare peridomestic, pale blue – Boyaca Peridomestic, green – Casanare sylvatic, pale green – SD. Human symbols indicate where clones from human cases are located on the tree. Red human symbols indicate co/superinfecting clones of likely sylvatic origin that cooccur in the same patient as clones from groups 14, 15 and 16. Arrows indicate congruence between the two clustering techniques and broadly define clades/clusters comprised of a majority of sylvatic (green) or domestic (red) clones.

© 2012 Blackwell Publishing Ltd
Peridomestic strains from BP (pale blue, NJ Tree, Fig. 1) and CP (pale blue, NJ Tree, Fig. 1) are poorly differentiated from those that occur in the local sylvatic environment as are a number of strains isolated from domestic vectors. The great majority of those clones isolated from humans (TcIDOM) are clustered in four populations 14, 15, 16 and 21. However, there are numerous instances where human clones from BD and SD are scattered across other populations linked to sylvatic and/or peridomestic transmission. Crucially, as indicated by red symbols on the NJ tree (Fig. 1), human patients are commonly infected with a mixture of clones, some belonging to the expected 14, 15, 16 and 21 groups, others highly divergent and nested among sylvatic clones (e.g. populations 2, 17 and 8).

Detection of mtDNA introgression events and mosaics

With the aim of identifying introgression events between nuclear and mitochondrial genomes, a second DAPC was conducted based on nuclear STR loci from the 100 isolates for which mtDNA sequences were available. In this case, 27 different PCs were retained and a total of 13 clusters identified among the data (Fig. 2). Group six corresponds to human clones from populations 14, 15 and 16 in the previous DAPC. These 13 groups were subsequently compared to patterns of clustering derived from maximum-likelihood composite (MLC) analysis of the mtDNA sequences (Fig. 2). Only MLC tree clades with >80% bootstrap support were treated as robust clusters. The mtDNA phylogeny supported the presence of significant diversity among

![Figure 2](image-url)
those isolates examined. Indeed, two highly divergent groups were present among the data: one apparently basal and closer to the TcII-TcIV outgroup and another, more populous, containing multiple well-defined subclades. Introgression was a pervasive phenomenon, and we identified multiple instances, among which the most striking are highlighted with red stars (Fig. 2). To rule out the possibility of ‘hidden’ or low frequency heteroplasmy across the two major mtDNA clades, we designed specific primers for ND1 and 9S sequences for the basal clade and attempted amplification of these sequences from key clones (AAC1c13, TmPA1c8, NR1c8, YAS1c13, primers in Table S2, Supporting information). No unexpected PCR products were observed.

In addition to detecting multiple signatures of mitochondrial introgression among clones, we uncovered a mosaic maxicircle sequence in LERcl11, FACS cloned from a patient in Santander province, using RDP v.3.0. Two parental clones were identified [LERcl12 (P1) and N5P14cl14 (P2), $P = 0.0374$]. Two recombination break points (break in and out) were found between at position 1196 (in) and 1379 (out), across a region spanning 14 SNPs. The mosaic was located in the MURF1 gene, within a continuous sequence fragment generated as part of the mMLST. To confirm the validity of the result, and in order to exclude PCR error and/or Taq polymerase template switching, we developed an ASP assay. Two distinct primer pairs were synthesized to amplify specifically each parent. Subsequently, forward P2 primer and reverse P1 primer were shown to amplify specifically LERcl11 but not parents in a touchdown PCR (Fig. S3; Table S2, Supporting information).

**BSP and TcI\textsubscript{DOM} genotype emergence estimation**

Both nuclear and mitochondrial data sets indicate the presence of a strongly bootstrap supported human domestic Tc genotype, TcI\textsubscript{DOM}, corresponding to TcIa (Herrera et al. 2007) or VEN\textsubscript{Dom} (Llewellyn et al. 2009). We used a Bayesian MCMC coalescent strategy to date the emergence of this genotype in the history of Tc in South America using maxicircle sequence data. Bayes factors identify expansion growth as the most appropriate demographic model under a strict molecular clock (log BF = 5.01, Table S5, Supporting information). The tMRCA of TcI\textsubscript{DOM} was estimated at 23 000 years (95% highest posterior density (HPD) 37 000–13 000 years) (ln $-15234.235$). A BSP corroborates this date and indicates a sharp reduction in effective population size approximately 27 000 years ago followed by a gradual expansion from 23 000 years ago until the present (Fig. 3).

**Discussion**

This study represents the most detailed attempt to document *Trypanosoma cruzi* population genetic diversity to date and combines high density spatiotemporal sampling, including multiple samples from the same host, with the application of high resolution genotyping of both nuclear and mitochondrial genomes to the resulting clones. A wealth of biologically, epidemiologically and evolutionarily important phenomena emerges.

---

**Fig. 3** Bayesian skyline plot estimating effective populations size fluctuations in the history of the domestic TcI genotype (TcIa/VEN\textsubscript{Dom}). Twenty million iterations were run to generate the highest credibility Bayesian tree based on the expansion growth model. Ninety-five per cent highest posterior density confidence intervals are presented, as well as a timeline.
Nuclear genetic diversity, population subdivision, superinfection and oral infection

Nuclear genetic diversity across the data set was uniformly high at all study sites and transmission cycles examined, with the exception of sylvatic isolates from Boyaca province. Repeated MLGs were surprisingly uncommon in the data set (57/269), given the number of samples that were clones from the same stock. Criscione & Blouin (2006) interpret high infrapopulation (within host) and intrapopulation (within study site) MLG diversity as evidence of low rates of clonal extinction at the population level among parasitic worms (Criscione & Blouin 2006). As with Plasmodium sp. (Schoepflin et al. 2009), we suggest this indicates intense local parasite transmission, consistent with extraordinary levels of Tcl diversity also present in the Venezuelan ‘llanos’ plains (Llewellyn et al. 2009, 2011), a continuous ecoregion linking western Venezuela and eastern Colombia that drains into the great Orinoco River. Using the same logic, reduced transmission and transmission efficiency should be accompanied by reduced parasite diversity. Sylvatic sites in Boyaca examined are heavily degraded and transformed by human activities. As such sylvatic mammal and triatomine capture success rates were significantly reduced. Deforestation and habitat destruction has been shown to significantly reduce parasite prevalence in sylvatic mammals in north-western Argentina (Ceballos et al. 2006). It seems a similar anthropogenic phenomenon could be negatively impacting sylvatic parasite diversity in Boyaca.

We chose to calculate population genetic statistics from clone corrected (identical MLGs deleted) groups of individuals assigned to populations a priori by study site and transmission cycle. Population assignment via DAPC and NJ D_AS clustering (adopted as a ‘model free’ approach, see Methods) indicates these groupings are in many cases genetically subdivided (Fig. 1). This is especially true for domestic and peridomestic populations, whereby a subset of clones is scattered among sylvatic parasite clades, while others group strongly together regardless of study site. Thus, widespread homozygosity in these populations might be attributed to Wahlund effects. Clones from CS form a more cohesive group in genetic terms. DAPC populations comprised of these isolates tend to cluster together, especially those making up group x in Fig. 1. Thus, we can perhaps give greater credence to significant homozygosity found therein. Several explanations compete to explain allelic homozygosity observed in trypansomone populations (Llewellyn et al. 2009; Rougeron et al. 2009). Most aim to account for the lack of a Meselson effect, which is thought to drive the emergence of extreme heterozygosity in the absence of sex over the long term (Mark Welch & Meselson 2000). From our reading of the literature, the Meselson effect seems to be an exception, rather than the rule for ‘clonal’ trypansomone population structures (Llewellyn et al. 2009; Rougeron et al. 2009; Gelanew et al. 2010; Barnabe et al. 2011; Kuhls et al. 2011; Llewellyn et al. 2011). In general, we urge caution in the interpretation of heterozygosity statistics at STR loci in the context of parasite sexuality, especially given that strong evidence for linkage disequilibrium often accompanies both negative and positive values for Fis. The current data set is no exception. For Tcl in particular, however, we note with interest that excess heterozygosity and low genetic diversity appear to go hand in hand at the population level (Llewellyn et al. 2009; Barnabe et al. 2011). This is also true for BS in our study, although we refrain from speculation as to the causes in the absence of more samples and better genomic coverage of the clones involved.

While nuclear STR loci alone in this study do little to enhance our understanding of parasite mating systems, patterns of clustering do assist source attribution of human and domestic T. cruzi clones. We have demonstrated that domestic and sylvatic Tcl populations were highly distinct in Venezuela and eastern Colombia (Herrera et al. 2007; Llewellyn et al. 2009) despite frequent invasion of the domestic setting by infected vector species. STR data presented here also support the existence of the same distinct domestic Tcl clade – TclDom. Comparisons between mitochondrial data sets support identity between this group and VENDom (human isolates from across Venezuela (Llewellyn et al. 2009; Messenger et al. 2012)) as well as Tcla (Herrera et al. 2007, p. 24). Previously, we interpreted sylvatic–domestic subdivision in the context of poor stercorarian transmission, whereby multiple exposures to contaminated domestic triatomine faeces were necessary to disperse the parasite, rather than opportunistic feeds by invasive sylvatic vectors (Llewellyn et al. 2009). Nonetheless, we were also able to demonstrate a low frequency of human symptomatic infection with sylvatic-type strains (Llewellyn et al. 2009). These new data clearly show that humans (e.g. EB, SEV and SP) can be infected simultaneously by both sylvatic and domestic-type T. cruzi clones. In regions highly endemic for Chagas disease, exposure commonly begins in early childhood. As with most other parasitic diseases, superinfection is thus a likely outcome and intrahost parasite diversity should accumulate with age, assuming a sufficient force of infection, incomplete cross-genotypic immunity and no density dependent factors influencing the ability of secondary infections to establish. No apparent relationship between age and infection multiplicity exists in our data set, and simultaneous
co-infection is also a likely source of intrahost parasite diversity. Furthermore, we have probably sampled only a proportion of intrahost diversity. In patients EM and EB, for example, mother and child respectively where congenital transmission has occurred, some MLGs are present in the child but not the mother. Two other points of interest relate to this congenital case. The first is that clones from the TcI_DOM can be transmitted congenitally, a phenomenon of wider importance throughout the region given its frequency among human cases (Herrera et al. 2007; Llewellyn et al. 2009; Cura et al. 2010). The second is that EM was coinfected with both TcI and TcII (GPI sequence identified), yet only TcI was isolated in EB. As such, our data suggest infection multiplicity cannot be ignored when evaluating the link between T. cruzi genotype and congenital transmission. Finally, we can confirm clones from several linked oral cases in Santander are likely to originate from the local sylvatic environment (patients EH, LER and SMA). However, domestic TcI clones within some of these individuals (LJVP and CACQ), suggest the foodstuff was contaminated by a mixed infection from a domestic vector(s). All cases were acute and symptomatic (patient EH died) inconsistent with the normal experimental outcome of T. cruzi superinfection and point to coinfection instead.

Mitochondrial introgression is not consistent with Trypanosoma cruzi clonality

Sexual activity in trypanosomes facilitates the emergence of successful epidemic clones and the spread of human infectivity in Trypanosoma brucei subspecies (Balmer et al. 2011), and new phenotypic traits in Leishmania (Volf et al. 2007). In T. cruzi, significant genetic exchange events that have shaped population structure – specifically those that have given rise to successful domestic DTUs TcV and TcVI – date to early human history (Lewis et al. 2011). It is known that T. cruzi possesses an extant capacity for (perhaps nonmeiotic) genetic exchange (Gaunt et al. 2003; Lewis et al. 2011). How the complex molecular toolkit for such a capacity might be retained, given its supposed infrequency of use, is a mystery. The solution presented herein is that TcI, at least, uses this machinery all the time. Figure 2 provides proof in the form of widespread incongruence between nuclear and mtDNA phylogenies, consistent with frequent genetic exchange between closely related, but distinguishable clones. Potentially common selfing events prevent a precise frequency estimate within our data set, however, examining Fig. 2, suggests a minimum of 17 events among 100 clones.

Introgression events are common among animal parasites. In both platyhelminthes [Schistosoma sp. (Huyse et al. 2009)] and nematodes [Ascaris sp., (Criscione et al. 2007)], mitochondrial introgression indicates contemporary hybridization between human and domestic animal species. Such events are evolutionarily advantageous in terms of expanding host ranges. Interestingly, as with our data, such events are frequently asymmetric, whereby little nuclear genetic exchange accompanies the invasive plastid genome (Steinauer et al. 2008). Asymmetric progeny from interspecies genetic crosses could result from chromosomal incompatibility between divergent parents (Lewis et al. 2011) or via chromosomal rearrangements between closely related strains. However, back crossing of hybrids into one parental population may also explain nuclear asymmetry in Schistosoma sp., as well as in T. cruzi. Similarly, we cannot rule out some nuclear exchange (Fig. S2, Supporting information), and while we have sampled perhaps 20% of the mitochondrial genome, we have sampled <0.1% of the nuclear genetic material from each clone.

Recombination among yeast, animal and plant mitochondrial genomes is widely reported (Rokas et al. 2004; Marechal & Brisson 2010; Solieri 2010). However, the maxicircle–maxicircle fusion and recombination detected and confirmed in our data set is a first among trypanosomes. Biparental inheritance of maxicircles is reported from experimental T. brucei crosses (Gibson et al. 2008). However, it is assumed that one population of maxicircles is lost stochastically in subsequent rounds of mitotic division. Patterns of maxicircle inheritance in our data set suggest an analogous mechanism pervades in 99% of cases. However, a 1% rate of maxicircle fusion translates into a significant number of potential recombinants population-wide. This result is hardly surprising given that mitochondrial genomes have a greater need than most to escape from Muller’s Ratchet because of their elevated mutation rate.

TcI_DOM population expansion coincides with the earliest human colonization of South America

Human colonization of South America is thought to have occurred via the Isthmus of Panama rapidly along the Pacific coastline before lateral migration eastwards into the interior. There are different reports of settlements in the late Pleistocene in South America and estimations of human arrival and settlement in South America fluctuate around 12 500 years ago (reviewed in Goebel et al. 2008). Remarkably, the colonization of North America via the Bering Land Bridge occurred a mere c. 2000 years earlier. The southerly migration of humans into South America coincides with the pattern of diversity we see in TcI_DOM, whereby nuclear and
mitochondrial markers suggest a link with sylvatic strains from Northern and Central America (Llewellyn et al. 2009; Cura et al. 2010; Messenger et al. 2012). The date we derive for the emergence of TcI (23 000 years, 95% HPD 37 000–13 000) corresponds broadly with the arrival of the earliest humans in the Americas and mirrors the level of uncertainty around the human molecular data [30 000–13 000 years ago (Goebel et al. 2008)]. Furthermore, the reduction in $N_e$ observed in this group, followed by a gradual expansion, fits with a pattern that might be expected during the colonization of a new host species/transmission cycle. Taken together, these observations suggest that early humans must have first domesticated this TcI genotype in North/Central America, prior to codispersion and coexpansion into South America. Indeed, SL-IR genotyping suggests a distribution that now extends as far south as the Argentine Chaco (Cura et al. 2010).

Conclusions

Next generation sequencing advances will shortly herald the arrival of the first protozoan parasite population genomic studies. However, this study demonstrates that classic population genetic studies can still be informative, especially where there is sufficient sample density in space, time, as well as from within each host or vector. Furthermore, we show that plastid vs. nuclear genome comparisons can provide important insight into the presence and mechanism of genetic exchange. Indeed, this data set provides probably the first ‘hard proof’ that genetic exchange in Trypanosoma cruzi occurs frequently in natural populations. Whole nuclear genomic comparisons will assist in identifying how asymmetrical such events truly are between parental strains. We hope these data will promote efforts to identify where genetic exchange occurs in the T. cruzi transmission cycle, as well as provide a model for identifying recombination among other cryptically sexual organisms.

Acknowledgements

Thanks to L. Rendón, E. Hincapíé, C. Herrera, M. Duque, A. Suárez and Y. Bogota for field and laboratory work support. Universidad de los Andes, the Wellcome Trust & EC FP7, Grant number 223034 (ChagasEpiNet) provided funds.

Author contributions

JDR wrote the paper, performed the research and analysed the data; FG wrote the paper; LAM analysed the data; MDL analysed the data; MM contributed materials; ZC contributed materials; MAM wrote the paper; and MSL designed the research, wrote the paper and analysed the data.

References


Barnabe C, De Meeus T, Noireau F et al. (2011) Trypanosoma cruzi discrete typing units (DTUs): microsatellite loci and population genetics of DTUs TcV and TcI in Bolivia and Peru. Infection, Genetics and Evolution, 11, 1752–1760.


Cura CI, Mejía-Jaramillo AM, Duffy T et al. (2010) Trypanosoma cruzi I genotypes in different geographical regions and transmission cycles based on a microsatellite motif of the intergenic spacer of spliced-leader genes. Experimental Parasitology, 118, 936–939.


Ramírez JD, Herrera C, Rendón L et al. (2011b) Estandarización de un protocolo de dilución limitante para la clonación celular de aislamientos colombianos de *Trypanosoma cruzi*. XX congreso Latinoamericano de Parasitología.


Ramírez JD, Herrera C, Rendón L et al. (2011b) Estandarización de un protocolo de dilución limitante para la clonación celular de aislamientos colombianos de *Trypanosoma cruzi*. XX congreso Latinoamericano de Parasitología.


Messenger L, Llewellyn M, Bhattacharyya T et al. (2012) Multiple mitochondrial introgression events and
J.D.R. and F.G. are fundamentally interested in the population genetics, phylogenetics, immunology and genomics of parasites. Currently they are working to elucidate the genetic structure of *Trypanosoma cruzi* and searching for plausible associations between the parasite's genetic variability and disease outcome. M.M. and Z.C. are interested in clinical aspects of Chagas Disease in Colombia. L.A.M., M.D.L., M.S.L. and M.A.M. specialise in the molecular epidemiology, pathology and genomics of trypanosomatid disease.

### Data accessibility

Sample distribution and metadata are included in the supporting information online (Table S1, Supporting information).

Microsatellite amplicon sizes are included in supporting information online (Table S4, Supporting information).

Sequence Data are deposited on Genbank accession numbers JQ585930–JQ586198 for GPI gene fragment, JX123135–JX123234 for the maxicircle sequences.

### Supporting information

Additional supporting information may be found in the online version of this article.

---

**Fig. S1** Distribution of 269 single-celled clones from Colombia.

**Fig. S2** Histogram showing the distribution of parent unique and shared alleles from seven selected TcI single-celled clones.

**Fig. S3** Electrophoresis gels results of allelic specific PCR assays confirming mosaic maxicircle sequence.

**Table S1** Location, habitat and host of *Trypanosoma cruzi* single-celled clones used in this study.

**Table S2** List of primers used in this study.

**Table S3.** Population specific genetic statistics.

**Table S4** Microsatellite allele sizes amplified at 24 loci across 269 clones.

**Table S5** BEAST estimates for the age of the domestic TcI clade.

Please note: Wiley-Blackwell are not responsible for the content or functionality of any supporting information supplied by the authors. Any queries (other than missing material) should be directed to the corresponding author for the article.
5.3.2. *Origins of natural hybrids in Colombia*

*T. cruzi* TcV and TcVI are known to be recent, natural inter-lineage hybrids, characterized by heterozygosity and minimal distinction, sharing intact alleles from their parental progenitors (TcII and TcIII) (Machado and Ayala, 2001; Brisse et al., 2003; Barnabé et al., 2011; Lewis et al., 2011; Yeo et al., 2011). The origin(s) of TcV and TcVI is unresolved; molecular dating indicates that these lineages evolved recently, within the last 60,000 years (Lewis et al., 2011), possibly from human disruption of sylvatic transmission cycles in the Southern Cone, suggesting there is continuous risk of genetic exchange driving the emergence of recombinant genotypes.

With increased sampling, the geographical range of TcV and TcVI appears to be significantly more extensive than previously suggested. On the basis of limited sequencing markers (*GPI*, cytochrome b and 18S rDNA), putative hybrid strains were identified among a minority of domestic and peridomestic strains from Colombia (Guhl and Ramírez, 2013), representing the first report of these lineages this far north; it remains unclear whether these are ‘bona fide’ TcV and TcVI strains or progeny of contemporary recombination events.

To resolve the putative status of these Colombian hybrids as novel recombinants, high resolution nuclear (Llewellyn et al., 2009a; Yeo et al., 2011; Diosque et al., 2014) and mitochondrial (Messenger et al., 2012) genotyping was used to characterize 24 novel Colombian strains (including 14 hybrid clones) in comparison to 33 additional reference isolates (TcII, TcIII, TcV and TcVI), from across South America.

This study is reported in full below in Messenger et al., submitted.

In summary:

- Based on MLMT data, all TcV and TcVI isolates were highly heterozygous (54.6% and 41.7% polymorphic loci with significant excess in heterozygosity, respectively) and displayed lower levels of genetic diversity (*D$_{AS}$* = 0.15 and 0.24; *A$_{r}$* = 2.38 and 2.46, respectively) and fewer private alleles (PA/L = 0.16 and 0.43, respectively) compared to TcII and TcIII parental strains (*D$_{AS}$* = 0.44 and 0.48; *A$_{r}$* = 3.94 and 4.26; PA/L = 1.76 and 2.35; 29.2% and 4.5% polymorphic loci with significant excess in heterozygosity, respectively). In this regard these lineages fulfil all of the expectations for progeny from a recent recombination event(s) with Mendelian allele inheritance.

- Haplotype resolution of five nuclear MLST targets demonstrated that, excluding infrequent incidences of LOH, each TcV and TcVI clone possessed intact TcII and TcIII alleles at every locus. Importantly, haplotypes in Colombian hybrids were either indistinguishable or closely related to those identified in reference TcVI strains, with minimal affinity to either TcII or TcIII Colombia-specific alleles.

- Mitochondrial MLST indicated that all Colombian hybrids had inherited a TcIII-type maxicircle, divergent from local TcIII mitochondria. A single Colombian isolate from a peridomestic dog, classified as TcVI by nMLST and MLMT, shared its maxicircle haplotype with southern TcV reference strains.
The majority of mitochondrial and microsatellite genotyping data confirmed that all 14 suspected Colombian hybrids were ‘bona fide’ TcVI clones, not novel recombinant strains, but were distinct from TcVI reference isolates from the Southern Cone.

Based on independent inheritance patterns of microsatellite loci, this dataset principally supported the hypothesis that two independent recombination events led to the formation of TcV and TcVI. However, a more parsimonious explanation for the sharing of mitochondrial haplotypes between Southern Cone TcV isolates and a Colombian TcVI strain may be that Colombian hybrids are a sibling group to their southern counterparts, which diverged and dispersed northwards, following a single hybridization event with biparental mitochondrial inheritance between heterozygous TcII and TcIII isolates.

Of the genotyping markers evaluated in this study, MLMT afforded the highest resolution, distinguishing all TcV and TcVI isolates and exposing intra-strain multiclonality among Colombian hybrids. mtMLST was also highly discriminatory, particularly among TcV isolates, but should not be used in the absence of comparative nuclear data. Among the five nMLST targets assessed, only TcAPX should be considered for future TcV and TcVI genotyping (DP=0.5 and 0.33, respectively).

The origin(s) of these Colombian TcVI strains remains unresolved. Additional sampling efforts are required to identify the primary vector species and explore local disease ecology, in order to assess the epidemiological risk of human Chagas disease associated with this hybrid lineage, considering its successful, epidemic establishment among domestic transmission cycles in the Southern Cone.
Please be aware that one cover sheet must be completed for each ‘Research Paper’ included in a thesis.

1. For a ‘research paper’ already published
   1.1. Where was the work published? ..............................................................................................................
   1.2. When was the work published? ....................................................................................................................
       1.2.1. If the work was published prior to registration for your research degree, give a brief rationale for its inclusion
                                                                                             ..................................................................................................
                                                                                             ..................................................................................................
   1.3. Was the work subject to academic peer review? .........................................................................................
   1.4. Have you retained the copyright for the work? Yes / No
       If yes, please attach evidence of retention.
       If no, or if the work is being included in its published format, please attach evidence of permission from copyright holder (publisher or other author) to include work

2. For a ‘research paper’ prepared for publication but not yet published
   2.1. Where is the work intended to be published? PLoS Neglected Tropical Diseases ..................................
   2.2. Please list the paper’s authors in the intended authorship order
       Messenger, L.A., Ramirez, J.D., Llewellyn, M.S., Guhl, F. & Miles, M.A
   2.3. Stage of publication – Not yet submitted [Submitted] Undergoing revision from peer reviewers’ comments / In press

3. For multi-authored work, give full details of your role in the research included in the paper and in the preparation of the paper. (Attach a further sheet if necessary)
   The candidate designed the study, performed all of the genotyping experiments, analyzed the data and drafted the manuscript.

NAME IN FULL (Block Capitals)  LOUISA ALEXANDRA MESSENGER

STUDENT ID NO: 223021

CANDIDATE’S SIGNATURE ____________________________ Date 10/2/15 

SUPERVISOR/SENIOR AUTHOR’S SIGNATURE (3 above)  

Improving health worldwide  www.lshtm.ac.uk
**Target Journal:** PLoS Neglected Tropical Diseases

**Title:** Origins of natural Trypanosoma cruzi hybrids in Colombia

Louisa A Messenger¹, Juan-David Ramirez², Martin S Llewellyn¹, Felipe Guhl² and Michael A Miles¹

¹Department of Pathogen Molecular Biology, Faculty of Infectious and Tropical Diseases, London School of Hygiene and Tropical Medicine, London, UK

²Unidad Clínico-Molecular de Enfermedades Infecciosas (UCMEI), Facultad de Ciencias Naturales y Matemáticas, Universidad del Rosario, Bogotá, Colombia

³Centro de Investigaciones en Microbiología y Parasitología Tropical (CIMPAT), Universidad de Los Andes, Bogotá, Colombia

**Abstract (250 words)**

**Keywords:** Chagas disease, Colombia, Trypanosoma cruzi, hybridization, TcV, TcVI

**Abstract**

**Background** The principal reproduction strategy of Trypanosoma cruzi, the aetiological agent of Chagas disease, is the subject of an intense, decades-old debate. Despite the existence of two recent natural hybrid lineages (TcV and TcVI), which are sympatric with severe human disease in southern endemic areas, a pervasive view is that recombination has been ‘restrained’ at an evolutionary scale and is of little epidemiological relevance to contemporary parasite populations. With increased sampling, the geographical range of TcV and TcVI is expanding, with putative hybrids identified as far north as Colombia.

**Methods** High resolution nuclear (25 microsatellite loci and 5 housekeeping gene fragments) and mitochondrial (10 loci) genotyping of hybrid clones from Colombia was undertaken, in comparison to representative reference strains from across South America, to resolve their putative status as novel recombinants.

**Results** All 14 suspected Colombian hybrids were highly heterozygous, minimally diverse and possessed intact parental alleles (TcII and TcIII) at each loci. Compared to local Colombian isolates, hybrid haplotypes were distinct from, but more closely related to, those identified in reference TcVI strains from the Southern Cone. Based on independent inheritance patterns of microsatellite loci, our data support the hypothesis that two separate recombination events led to the formation of TcV and TcVI. However, more private alleles among Colombian hybrids and the sharing of mitochondrial haplotypes between southern TcV isolates and a Colombian TcVI strain, suggests the evolution of these recombinant lineages may be more complicated than previously assumed.

**Conclusions** The origin of these Colombian hybrids is unclear; they are unlikely to be predecessors of southern TcVI strains, but are also not clear descendants, and may instead
represent a sibling group, which diverged and dispersed northwards, following a single hybridization event between heterozygous TcII and TcIII isolates in the Southern Cone. Importantly, the geographical range expansion of TcVI has potential implications for human Chagas disease in Colombia, considering the successful, epidemic establishment of this lineage among domestic transmission cycles in the Southern Cone.

Introduction

Many eukaryotic pathogenic microorganisms (both fungal and protozoan) that were previously assumed to reproduce clonally have retained non-obligate, cryptic sexual cycles (Heitman, 2010). Genetic exchange has the potential to drive the evolution of novel recombinant strains with epidemiologically significant traits, including increased pathogenicity, transmissibility and drug resistance (Awadalla, 2003). However, limiting sexual reproduction allows the generation of host-adapted clonal populations that retain the ability to hybridize in response to selective pressures. The conservation of meiotic gene orthologues among several basally divergent protists, including Giardia (Poxleitner et al., 2008), Entamoeba (Ramesh et al., 2005) and Trichomonas (Malik et al., 2008) suggests that the common ancestor of all eukaryotes was capable of meiotic recombination.

Chagas disease is the most important vector-borne parasitic infection in Latin America, affecting an estimated 8-10 million individuals, with a further 90 million at risk (Hotez et al., 2008). The aetiological agent, Trypanosoma cruzi (Kinetoplastida: Trypanosomatidae), is a complex vector-borne zoonosis, transmitted by more than 100 species of hematophagous triatomine bugs (Hemiptera: Reduviidae: Triatominae) (Lent and Wygodzinsky, 1979; Galvão et al., 2003) and maintained by mammalian reservoir hosts ranging from the southern United States to Argentinean Patagonia (Noireau et al., 2009). Human Chagas disease is principally restricted to areas where individuals are exposed to infected vector faeces through contact with intact mucosae or abraded skin (Coura and Dias, 2009). In the absence of chemotherapy, the majority of infected individuals are asymptomatic for life. However, over a period of 10-30 years, approximately 20-30% will develop irreversible, potentially fatal cardiac syndromes, or more rarely, dilatation of the gastrointestinal tract (megaoesophagus or megacolon) (Rassi Jr et al., 2010).

T. cruzi displays remarkable genetic diversity, which has long been considered a principal factor underlying the major clinical variation observed in Chagas disease (Miles et al., 2009). Current international consensus recognizes a minimum of six stable genetic lineages or discrete typing units (DTUs): TcI-TcVI (Zingales et al., 2009). Molecular analyses indicate that T. cruzi has a predominantly clonal population structure, interspersed with infrequent genetic exchange events. DTUs TcI-TcIV form monophyletic clades and TcV and TcVI are known to be recent natural inter-lineage hybrids (Machado and Ayala, 2001; Lewis et al., 2011). Multilocus sequence typing (MLST) supports these designations with TcI-TcIV characterized by substantial allelic homozygosity, likely resulting from recurrent, genome-wide and dispersed gene conversion. TcV and TcVI display natural heterozygosity and
minimal distinction, sharing intact alleles from their parental progenitors (TcII and TcIII),
(Machado and Ayala, 2001; Brisse et al., 2003; Barnabè et al., 2011; Lewis et al., 2011; Yeo
et al., 2011).

The origin(s) of TcV and TcVI is unresolved and it is presently contested whether they arose
from two independent genetic exchange events (de Freitas et al., 2006; Lewis et al., 2011) or
a single incidence of hybridization followed by clonal divergence (Westenberger et al., 2005;
Sturm and Campbell, 2010; Flores-López and Machado, 2011). Molecular dating indicates
that these lineages evolved recently, within the last 60,000 years (Lewis et al., 2011),
possibly from human disruption of sylvatic transmission cycles in the Southern Cone,
suggesting there is continuous risk of genetic exchange driving the emergence of novel
recombinants (Flores-López and Machado, 2011; Lewis et al., 2011).

Currently, the frequency of natural recombination in T. cruzi is still debated, as are the
underlying cytological mechanisms (Tibayrenc and Ayala, 2012; 2013; Ramírez and
Llewellyn, 2014). Characterization of experimental intra-TcI hybrids suggests that
hybridization may occur within the mammalian host and is analogous to the parasexual cycle
of Candida albicans; nuclear fusion creates a tetraploid intermediate, followed by
homologous recombination, gradual genome erosion and reversion to aneuploidy (Gaunt et
al., 2003; Lewis et al., 2009). However, this unusual mating system differs from canonical
meiosis and is challenging to reconcile with both the presence of highly conserved meiotic
orthologues within the T. cruzi genome (Ramesh et al., 2005) and the existence of natural
diploid heterozygous lineages, which resemble Mendelian F1 progeny (TcV and TcVI)
(Lewis et al., 2009; 2011).

Historically, most T. cruzi DTUs have had broadly distinct, but often overlapping,
geographical and ecological distributions (Miles et al., 2009; Zingales et al., 2012). In
general, TcI, TcII, TcV and TcVI are frequently isolated from domestic cycles and are
responsible for the majority of human infections. TcI is the principal cause of human Chagas
disease in northern South America (Anez et al., 2004; Ramírez et al., 2010; Carrasco et al.,
2012). By comparison, TcII, TcV and TcVI are largely confined to domestic transmission
cycles in southern parts of South America, overlapping with severe human Chagas disease,
(Corrales et al., 2009; Burgos et al., 2010; Bisio et al., 2011). The sylvatic reservoirs of these
three DTUs are not fully defined, although TcII has been increasingly isolated from primates
in Brazil (Fernandes et al., 1999; Lisboa et al., 2007; Araújo et al., 2011); peridomestic dogs
are emerging as potential hosts of TcV and TcVI (Maffey et al., 2012; Enriquez et al., 2013;
Fernandez et al., 2014).

With increased sampling, the geographical range of TcV and TcVI appears to be significantly
more extensive than previously suggested. On the basis of limited sequencing markers (GPI,
cytochrome b and 18S rDNA), putative hybrid strains were identified among a minority of
domestic and peridomestic strains from Colombia (Guhl and Ramírez, 2013), representing
the first report of these lineages this far north; it remains unclear whether these are ‘bona
fide’ TcV and TcVI strains or progeny of a novel recombination event. Elucidating the
molecular epidemiology of TcV and TcVI is complicated by limited sample collections and difficulties distinguishing these strains from each other (Venegas et al., 2011) or even their parental lineages (TcII and TcIII) (Yeo et al., 2011; Diosque et al., 2014). However, an improved understanding of the frequency of genetic exchange in T. cruzi and the impact of hybridization on parasite genetic diversity, natural population structures and their ecological and geographical distributions, are crucial to provide an accurate assessment of the epidemiological risk associated with recombinant genotypes.

In this study we performed high resolution nuclear and mitochondrial genotyping of novel natural hybrid clones from Colombia in comparison to reference strains from across South America to resolve their putative status as novel recombinants. Additionally an in-depth analysis of allele inheritance of three different type of molecular markers was undertaken to characterize intra-TcV and TcVI genetic diversity across their expanding geographical range.

**Materials and methods**

**Parasite strains and DTU assignment**

A panel of 57 T. cruzi biological clones was assembled for analysis, including 24 uncharacterized strains from North Colombia and 33 additional reference clones (TcII, TcIII, TcV and TcVI) (Table 1; Figure 1). Colombian strains were isolated from human patients (adults and suspected congenitally-infected infants), triatomine bugs (Panstrongylus geniculatus, Rhodnius prolixus and Triatoma venosa) and sylvatic mammalian hosts (Dasypus species). Supporting reference clones were derived from a range of hosts and vectors from locations representative of both intra-lineage genetic diversity and geographical distribution of each DTU (Figure 1).

Biological clones were obtained from primary cultures by either fluorescence-activated cell sorting (Valadares et al., 2012) or plate cloning (Yeo et al., 2007). Colombian isolates were initially assigned to DTU-level via PCR amplification of the SL-IR, 24Sa rDNA and 18S rDNA subunits according to Guhl and Ramírez, 2013. Putative hybrid strains were identified by either a double 24Sa rDNA amplicon (125 and 140 bp) (TcV) or single 24Sa rDNA amplicon (140 bp) and amplification of the A10 fragment of the 18S rDNA subunit (TcVI) (525 or 630 bp), and confirmed by sequencing glucose-6-phosphate isomerase (GPI), as previously described (Guhl and Ramírez, 2013).

**Nuclear Multilocus Sequence Typing (nMLST)**

Five nuclear housekeeping genes (glutathione peroxidase GPX; GTP-binding protein RAB7 GTP; metacyclin-II Met-II; ascorbate-dependent haemoperoxidase TcAPX; and mitochondrial peroxidase TcMPX), selected from two multilocus sequence typing schemes on the basis of
their intra-TcV and TcVI discriminatory powers (Yeo et al. 2011; Lauthier et al. 2012), were amplified and sequenced across all samples according to Yeo et al. 2011.

Nucleotide sequences were assembled manually in BioEdit v7.1.3.0 sequence alignment editor software (Ibis Biosciences, USA) (Hall, 1999) and unambiguous consensus sequences were produced for each hetero. Heterozygous SNPs were visually identified by the presence of two coincident peaks at the same locus (‘split peaks’), verified in forward and reverse sequences, and manually scored according to the one-letter nomenclature for nucleotides from the International Union of Pure and Applied Chemistry (IUPAC).

Initially, diploid sequence data were analysed per locus in MLSTest (Tomasini et al. 2013); heterozygous single nucleotide polymorphisms (SNPs) were handled as average states. Individual Neighbour-Joining (NJ) trees were constructed for each gene and the extent of phylogenetic incongruence between targets was evaluated visually. Isolates were classified into DTUs and sequence types (ST), discriminatory powers (DP) and typing efficiencies (TE) were calculated per locus for the complete dataset and separately for Colombian strains (Table 2).

For each isolate, diploid sequence data were then concatenated in order of their relative chromosomal positions (Met-II, GTP, TcMPX, TcGPX and TcAPX, on chromosomes 6, 12, 22, 35 and 36, respectively). In MLSTest, phylogenetic incongruence between loci was assessed using the BIO-Neighbour Joining Incongruence Length Difference test (BIONJ-ILD) and evaluated by a permutation test with 1000 replicates. A final NJ tree was constructed and statistical support was evaluated by 1000 bootstrap replications. Incongruent branches in the concatenated NJ tree were identified using the Neighbour-Joining Localized Incongruence Length Difference (NJ-LILD) test.

To examine patterns of TcII and TcIII allele inheritance in TcV and TcVI isolates, haplotypes for each gene were inferred using PHASE software version 2.1 (Stephens et al., 2001) implemented in DnaSP v5.10.1 (Librado and Rozas, 2009). For isolates where PHASE was unable to adequately resolve haplotypes (uncertainty probability p<0.95), PCR products were cloned and sequenced to experimentally confirm predicted gene phase. PCR products were cloned using the pGEM®-T Easy Vector system I (Promega, UK), according to the manufacturer’s instructions. Plasmids were sequenced from a minimum of six colonies per isolate using standard T Easy Vector primers T7 (5’-TAATACGACTCACTATAGGG-3’) and Sp6 (5’-ATTAGGTGACACTATAG-3’) (Promega, UK). In cases where haplotypes remained ambiguous, a further six colonies were picked and processed, as described.

Following haplotype resolution, the most appropriate nucleotide substitution model for each gene was selected from 1,624 candidates, based on the Akaike Information Criterion (AIC), in jMODELTEST 2.1.4 (Darriba et al., 2012) and used to construct individual Maximum-Likelihood (ML) phylogenies in MEGA 6 (Tamura et al., 2013). Bootstrap support for clade topologies was estimated following the generation of 1000 pseudo-replicate datasets. Bayesian phylogenetic analysis was performed per loci with MrBAYES, implemented through TOPALi v2.5, using the best-fit model, based on the Bayesian Information Criterion.
(BIC) (Milne et al., 2009). Five independent analyses were run for one million generations, with sampling every 100 simulations (25% burn-in). Statistically-supported topological incongruence between individual ML and Bayesian phylogenies was evaluated using Kishino-Hasegawa (KH) (Kishino and Hasegawa, 1989) likelihood tests in PAML v.4 (Yang, 2007).

**Multilocus Microsatellite Typing (MLMT)**

Twenty-five nuclear microsatellite loci were amplified as previously described by Llewellyn et al., 2009. These markers are distributed across eleven putative chromosomes, including five groups of physically linked loci (Weatherly et al., 2009). A full list of microsatellite targets and primers are given in Table S2. Allele sizes were determined using an automated capillary sequencer (AB3730, Applied Biosystems, UK), in conjunction with a fluorescently tagged size standard (GeneScan™ – 500 LIZ®, Applied Biosystems, UK), and manually checked for errors in GeneMapper® software v3.7. All isolates were typed ‘blind’ to control for user bias (Supplementary file S1).

Individual-level sample clustering was initially defined using a Neighbour-Joining (NJ) tree based on pair-wise distances ($D_{AS}$: 1 – proportion of shared alleles at all loci/n) between microsatellite genotypes calculated in MICROSAT v1.5d (Minch et al., 1997) under the infinite-alleles model (IAM). To accommodate multi-allelic genotypes ($\geq 3$ alleles per locus), a script was written in Microsoft Visual Basic to generate random multiple diploid resamplings of each multilocus profile. A final pair-wise distance matrix was derived from the mean across multiple re-sampled datasets and used to construct a NJ phylogenetic tree in PHYLIP v3.67 (Felsenstein, 1989). Majority rule consensus analysis of 10,000 bootstrap trees was performed in PHYLIP v3.67 by combining 100 bootstrap samples generated in MICROSAT v1.5d, each drawn from 100 respective randomly re-sampled datasets.

A single randomly sampled diploid dataset was used for all subsequent analyses. DTU-level genetic diversity was evaluated using sample size corrected allelic richness ($A_r$) in FSTAT 2.9.3.2 (Goudet, 1995). Sample size corrected private (lineage-specific) allele frequency per locus (PA/L) was calculated in HP-Rare (Kalinowski, 2005). DTU-level heterozygosity indices were calculated in ARLEQUIN v3.11 and associated significance levels for p-values derived after performing a sequential Bonferroni correction to minimise the likelihood of Type 1 errors (Rice, 1989). All population genetics parameters were derived for both the complete dataset and separately for Colombian strains (Table 3).

Microsatellite allele inheritance among all TcV and TcVI isolates was examined manually. At each locus genotypes were classified as hybrid (TcII/TcIII), or non-hybrid (TcII/TcII or TcII/TcIII) based on presence or absence of TcII- and TcIII-specific alleles among parental strains. A separate analysis was performed for Colombian hybrid clones, based on the identification of Colombia-specific parental alleles.
Mitochondrial Multilocus Sequence Typing (mtMLST)

For all isolates, ten mitochondrial gene fragments were amplified, sequenced and concatenated according to Messenger et al., 2012 and are available from GenBank under the accession numbers listed in Table S1. The most appropriate nucleotide substitution model was selected from 1,624 candidates, based on the AIC, in jMODELTEST 2.1.4 (Darriba et al., 2012). Alternate Maximum-Likelihood (ML) phylogenies were constructed using the GTR+G model (nine substitution rate categories) in MEGA 6 (Tamura et al., 2013). Bootstrap support for clade topologies was estimated following the generation of 1000 pseudo-replicate datasets. Bayesian phylogenetic analysis was performed with MrBAYES, implemented through TOPALi v2.5, using the best-fit model, based on the Bayesian Information Criterion (BIC) (HKY+G) (Milne et al., 2009). Five independent analyses were run for one million generations, with sampling every 100 simulations (25% burn-in). Using the concatenated dataset (excluding indels), STs, DPs and TEs were calculated for all isolates and separately for Colombia strains (Table 2).

Results

nMLST

Diploid sequence data were concatenated across all five loci to produce a gap free alignment of 2439 bp. A total of 44 unique STs and 74 variable sites (VSs) were identified (~3.03% sequence diversity). All TcV and TcVI isolates consistently displayed heterozygous profiles. Complete loss of heterozygosity (LOH) was observed among a subset of hybrid strains at four loci: within GPX for P251 cl7 (loss of TcII-type allele), GTP for AACf2 cl11 (loss of TcIII-type allele), Met-II for Para6 cl4 (loss of TcIII-type allele) and TcMPX for 7/8 TcV isolates (excluding PAH179 cl5) (loss of TcIII-type allele) and two TcVI clones (LHVA cl4 and P251 cl7) (loss of TcII-type allele). Partial LOH was observed within Met-II for EPV20-1 cl11 (5’ loss of TcII-type allele).

Individual genes varied with respect to sequence diversity, ranging from 7.2% for Met-II to 1.58% for TcMPX. Number of sequence types (STs), typing efficiency (TE) and discriminatory power (DP) per locus and DTU are given in Table 2. Of the five MLST targets evaluated, GPX, Met-II and TcAPX were the most discriminatory (DP = 0.3, 0.33 and 0.35, respectively), identifying 17, 19 and 20 unique STs (from 57 clones), respectively. At the intra-DTU level these three loci were also highly resolutive (Table 2), in particular, for TcV and TcVI, TcAPX distinguished the most STs (DP = 0.5 and 0.33, respectively).

In general, TcV and TcVI were characterized by the lowest levels of intra-lineage genetic diversity; for the former, two out of five genes (GPX and GTP) were monomorphic. When considering Colombian strains separately, TcII isolates were the most homogeneous (identical across all five MLST loci). Intra-strain clonal diversity was generally low, but detected among some Colombian TcVI strains (2/2, 3/5 and 2/4 unique genotypes for strains PG98, Rp540 and VS, respectively). As expected the concatenated dataset afforded higher
intra-lineage resolution than individual loci, discriminating the most unique intra-DTU STs (10/15 TcII, 13/13 TcIII, 5/8 TcV and 16/21 TcVI) (Table 2).

A Neighbour-Joining phylogeny constructed from concatenated sequence data (Figure 2A) robustly identified TcII (99%) and TcIII (100%) as monophyletic clades. All TcV isolates, with the exception of Para6 cl4, grouped together (70%), while TcVI strains were clustered basally to TcII and TcV with poorly-supported internal branching. Moderate statistically-supported incongruence was detected between loci (BIOMJ-ILD p<0.001) and localized among TcV and TcVI branch nodes, likely attributable to irregular LOH, and accounting for the failure to recover these DTUs as discrete clades within the concatenated tree.

Following in silico and experimental haplotype resolution, Maximum-Likelihood and Bayesian phylogenies were generated for each locus individually (an example is given in Figure 2B). In all cases, excluding the aforementioned instances of LOH, TcV and TcVI clones each possessed one TcII- and one TcIII-derived haplotype. In Figure 2B, haplotypes for Met-II were robustly clustered into two major clades corresponding to TcII (94/0.98) and TcIII (99/1.0). For this locus all TcV and TcVI strains possessed a TcII-like haplotype, which was unique to each DTU and distinct from other TcII reference clones, but shared their TcIII allele with geographically-disparate TcIII reference isolates. Other variations of allele inheritance were observed across the other four genes, for example, for GTP and TcGPX, TcII- and TcIII-type haplotypes were shared across both hybrid DTUs (data not shown).

With regards to the geographical origin of hybrid TcII and TcIII alleles, for some genes insufficient genetic diversity was present (GTP, TcMPX) to derive any correlation, while others (GPX, Met-II and TcAPX) supported a putative association of hybrid TcIII-type haplotypes with those observed in southern TcIII clones from Bolivia, Peru and Paraguay. Importantly, at all five loci, the majority of Colombian haplotypes were either indistinguishable from or closely related to those identified in reference TcVI strains, with minimal affinity to either TcII or TcIII Colombia-specific alleles.

MLMT

Twenty-five microsatellite loci afforded the highest resolution of any type of marker; all isolates were characterized by unique multilocus genotypes (MLGs), with the exception of EB cl4 and cl6 (Figure 3 and Table 3). Based on both nMLST and MLMT datasets, all putative Colombian hybrids were classified as TcVI (Table 1).

Consistent with nMLST data, TcV and TcVI displayed lower levels of genetic diversity ($D_{AS} = 0.15$ and $0.24$; $A_r = 2.38$ and 2.46, respectively), private alleles (PA/L = 0.16 and 0.43) and heterozygosity (54.6% and 41.7% polymorphic loci with significant excess in heterozygosity, respectively) compared to their parental lineages ($D_{AS} = 0.44$ and 0.48; $A_r = 3.94$ and 4.26; PA/L = 1.76 and 2.35; 29.2% and 4.5% polymorphic loci with significant excess in heterozygosity for TcII and TcIII, respectively), supporting their hybrid status (Table 3). However, when considering only Colombian isolates, TcII clones emerged as the least diverse ($D_{AS} = 0.062$; $A_r = 1.65$), in agreement with the nMLST data. Colombian clones also
possessed more private alleles per locus but lower levels of allelic richness compared to reference TcVI strains (PA/L = 0.86; Ar = 1.87 and PA/L = 0.43; Ar = 2.46, respectively).

A NJ tree based on $D_{AS}$ values robustly separated each DTU (Figure 3). The TcIII clade was strongly-subdivided (99%), corresponding to a northern (Brazil and Colombia) and a southern (Bolivia, Paraguay and Peru) group. Substructuring was also evident among TcII clones; Colombian isolates were grouped apart from all other reference strains (83%). By comparison, within the TcVI clade, all Colombian clones were unique but most were interspersed among reference strains; a subset of domestic isolates was strongly separated (96%).

Patterns of TcII and TcIII allele inheritance were examined for all isolates and specifically for Colombian hybrids. Alleles were classified as TcII-specific, TcIII-specific, shared between TcII and TcIII or private to either TcV or TcVI. Most loci were heterozygous for either a TcII allele and a TcIII allele, one parental allele and one shared allele or two shared alleles; a minority were homozygous for one of either parental alleles or one shared allele. Fixed inter-lineage genotypic differences were observed at 84% (21/25) of microsatellite loci and of the alleles that distinguished between hybrid DTUs, 70.4% (38/54) were shared by parental strains.

For TcV isolates, 25.1% of alleles were TcII-specific, 29.9% TcIII-specific, 39.8% shared between TcII and TcIII and 5.08% private. For TcVI strains, 18.3% of alleles were TcII-specific, 29.5% TcIII-specific, 41.9% shared between TcII and TcIII and 10.2% private.

Considering only Colombian hybrids, 14.4% of alleles were TcII-specific, 27% TcIII-specific, 42.9% shared between TcII and TcIII, 1.45% Colombian TcII-specific, 1.74% Colombian TcIII-specific and 12.5% private.

**mtMLST**

Ten maxicircle gene fragments were sequenced across all strains and concatenated to produce a 3647 bp alignment; no phylogenetic incongruence was observed between individual loci. Thirty-six unique haplotypes were identified from a total of 774 variable sites (~21.2% sequence diversity). Numerous small indels (1-3 bp) were identified, as well as a large deletion of 245 bp within ND1, shared among 5/15 TcII clones (3/6 Colombian), and within MURF1 fragment b, a 17 bp deletion common to the same TcII strains, and a 27 bp insertion unique to CM25 cl2. By comparison to TcIII, intra-lineage diversity in TcII, TcV and TcVI was markedly lower (Table 2).

Maximum-Likelihood and Bayesian phylogenies were constructed from the concatenated maxicircle data and were not significantly incongruent (Figure 4). Isolates were grouped into two monophyletic clades, corresponding to TcII (100/1.0) and TcIII+TcV+TcVI (100/1.0).

Within the TcII clade, Colombian clones were separated into two strongly-supported subgroups; no clear correlation by host or geography was identified. Human Colombian TcII isolates clustered with a subset of domestic strains (both from humans and *T. infestans*) from...
Brazil and Paraguay (99/1.0), while domestic triatomine clones were grouped with others from Chile and Bolivia (100/1.0).

Robust internal branching was also observed within the TcIII+TcV+TcVI clade; clones from each DTU were clustered together in subclades. All TcVI strains were grouped together, however, Colombian maxicircle haplotypes were diverse between study sites and noticeably distinct from reference TcVI isolates. Interestingly, AACf2 cl11, a new peridomestic Colombian isolate from a dog, which was unequivocally classified as a TcVI by nMLST and MLMT, appeared to possess a TcV-type maxicircle haplotype (Figure 4). Both TcV and TcVI mitochondrial haplotypes were most closely related to TcIII-type maxicircles found in domestic/peridomestic strains from Peru and Paraguay (SABP19 cl1 and X9/3 and X109/2, respectively).

Discussion

This study exploited high resolution nuclear and mitochondrial genotyping to establish whether putative hybrid strains from Colombia were the progeny of a novel recombination event or, more likely, an extension of lineage geographical range. Additionally, we undertook a detailed analysis of intra-TcV and TcVI genetic diversity, at overlapping levels of resolution, to further characterize two of the most poorly described T. cruzi DTUs.

TcV and TcVI intra-lineage genetic diversity

Consistent with previous reports (Machado and Ayala, 2001; Brisse et al., 2003; Westenberger et al., 2005; Lewis et al., 2011; Yeo et al., 2011), TcV and TcVI isolates were heterozygous, possessed intact parental alleles at an approximate 1:1 ratio, and displayed lower levels of genetic diversity and fewer private alleles/SNPs compared to parental DTUs; in this regard these lineages fulfilled all of the expectations of progeny from a recent recombination event(s) with Mendelian allele inheritance. The minority of homozygous loci detected by nMLST and MLMT can be explained by random LOH resulting from gene conversion, which has been described in T. cruzi at varying frequencies (Yeo et al., 2011; Barnabé et al., 2011; Diosque et al., 2014). These observations caution reliance on a single locus for DTU assignment of suspected hybrids, particularly in areas where these lineages are sympatric with their parental genotypes.

The failure by others to separate TcV and TcVI isolates likely reflects the number and discriminatory power of different genotyping targets used (de Freitas et al., 2006; Venegas et al., 2011; Barnabé and Breniere, 2012; Perez et al., 2013). To our knowledge, this study employed the most loci to date, exposing intra-strain multiclonality among Colombian hybrids, suggesting there is additional, un-sampled genetic diversity hidden at the hybrid intra-lineage level. Of the markers assessed, the MLMT scheme afforded the highest resolution, discriminating between all TcV and TcVI isolates. Individual housekeeping genes were unable to distinguish most hybrid strains; of the five under evaluation, only TeAPX warrants consideration for future characterization of TcV/TcVI DTUs. The concatenated mitochondrial dataset was more resolutive, enabling the identification of all TeV strains, but...
should not be used in the absence of comparative nuclear data as, in agreement with previous studies, all hybrid strains had inherited a TcIII-type maxicircle haplotype (Machado and Ayala, 2001; Westenberger et al., 2005; de Freitas et al., 2006; Lewis et al., 2011; Barnabé and Breniere, 2012).

Evolutionary origins of Colombian hybrids and TcV and TcVI

The majority of genotyping data confirmed that all 14 suspected Colombian hybrids were ‘bona fide’ TcVI clones and not novel recombinant strains. Examination of nuclear haplotypes across five loci demonstrated that Colombian hybrid TcII and TcIII alleles were shared amongst other TcVI strains from the Southern Cone and not closely related to unique parental alleles from Colombia. MLMT also supported this pattern of inheritance, with only a minority of Colombian private parental alleles shared by hybrids from the same area. A number of Colombian TcII- (nine alleles among eight loci) and TcIII (14 alleles among 11 loci) -specific alleles were identified, and most of these differed from reference alleles by up to 15 bp, suggesting homoplasy was unlikely to hinder our detection of any parental contributions to hybrids. By mitochondrial loci, Colombian TcVI isolates were also divergent from local TcIII maxicircle haplotypes.

All three genotyping schemes demonstrated that Colombian TcVI clones were related to, but distinct from, TcVI reference strains from the Southern Cone, raising the question of where these hybrids originated from? Colombian TcVI clones had more private microsatellite alleles compared to their southern counterparts, tentatively suggesting they may be ancestral. This is not supported by allelic richness measurements among these hybrids, which were lower than other TcVI clones, although this might be attributable to sampling bias; multiple clones per Colombian strain (between two to five) from a restricted study area were examined vs. single references clones from across a much wider geographical range.

Currently, the origins of TcV and TcVI are the subject of a popular, on-going debate. Based on multiple nuclear and mitochondrial sequencing markers, it has been suggested that these DTUs were the product of two independent genetic exchange events between TcII and TcIII (de Freitas et al., 2006; Lewis et al., 2011); investigation of only nuclear loci supports a scenario where a single incidence of hybridization was followed by clonal divergence (Westenberger et al., 2005; Flores-López and Machado, 2011). Overall, our data concur with the former hypothesis; TcV and TcVI have distinct nuclear and mitochondrial MLST genotypes, related but independent microsatellite allelic profiles and the majority of alleles that distinguish between hybrid DTUs were also shared by parental strains. If inter-lineage differences were the result of clonal divergence, at rapidly evolving microsatellite loci, a much higher frequency of private alleles would be expected. However, the identification of a TcV-type maxicircle within a TcVI Colombian strain (AACf2 cl11) introduces a slight degree of uncertainty; all isolates in this study were biological clones, ruling out mixed infections as possible confounders.
Considering Colombian TcVI clones were unlikely to be predecessors of TcVI strains from the Southern Cone, where there is compelling evidence to suggest hybrids and parents are of local origin (Westenberger et al., 2006; Lewis et al., 2011), but were also not clear descendants, a more parsimonious explanation might be that they are a sibling group.

Assuming *T. cruzi* mitochondrial inheritance, under exceptional (or perhaps conventional) circumstances can be biparental (as observed in Colombian TcI isolates (Ramirez et al., 2012)), and that TcV and TcVI evolved from the beneficiaries of different alleles during a single hybridization event between heterozygous parents, then the presence of a TcV-type maxicircle in a Colombian TcVI clone may simply be the result of incomplete lineage sorting. An indistinguishable pattern of inheritance would also be observed following a recent mitochondrial introgression event from TcV into TcVI which either left undetectable signatures of nuclear hybridization by our markers, or possibly none at all (Messenger et al., 2012; Ramirez et al., 2012). However, genetic exchange has not been described in hybrid DTUs previously, but might be expected to be more permissive between closely related strains (Ramirez and Llewellyn, 2014). Nor has TcV been unequivocally identified in Colombia, given the single putative isolate (AACf2 cl11) identified previously (Guhl and Ramirez, 2013) was genotyped in this study as a TcVI. It is noteworthy that AACf2 cl11 was isolated from a peridomestic dog; genetic exchange in *T. cruzi* has been proposed to arise within mammalian cells (Gaunt et al., 2003) and mixed infections in such hosts are a common (Crisante et al., 2006; Ramirez et al. 2013a), potentially provide opportunities for recombination to occur. Without additional hybrid samples from northern Brazil and Venezuela, it is impossible to confirm either hypothesis or infer the directionality of hybrid dispersal.

**Implications for human Chagas disease in Colombia**

With parallel improvements in sampling strategies and genotyping techniques, our understanding of the geographical and ecological distribution of each *T. cruzi* DTU is changing. Currently, in Colombia, human Chagas disease is principally associated with TcI (Ramirez et al., 2010; Guhl and Ramirez, 2013), to a lesser extent TcII (Zafra et al., 2008; Mantilla et al., 2010; Ramirez et al., 2010), and TcIV during microepidemic oral outbreaks (Ramirez et al., 2013b). While hybrid infections can be more challenging to distinguish from their parental progenitors, TcVI does not thus far appear to play a prominent role in local disease transmission.

In this study, Colombian hybrids were isolated from *P. geniculatus, R. prolixus* and from two human infections, one associated with congenital transmission, the other more likely vector-borne. Further intensive sampling efforts are required to identify the primary domestic vector species and transmission cycle ecology of TcVI in Colombia, in order to accurately assess the epidemiological risk of human Chagas disease associated with this hybrid lineage, considering its successful, epidemic establishment among domestic transmission cycles in the Southern Cone.
References


Figure 1. Map showing geographical distributions of TcII, TcIII, TcV and TcVI clones.

A panel of 57 T. cruzi biological clones was assembled for analysis, including 24 isolates from North Colombia and 33 additional reference strains, derived from a range of hosts and vectors representative of both intra-lineage genetic diversity and geographical distribution of each DTU. Colombian strains were isolated from human patients (adults and suspected congenitally-infected infants), triatomine bugs (Panstrongylus geniculatus, Rhodnius prolixus and Triatoma venosa) and sylvatic mammalian hosts (Dasypus species). Closed circles indicate origin of biological clones and colours denote isolate DTU (TcII, TcIII, TcV or TcVI) (see legend).
Figure 2. A: Unrooted Neighbour-Joining tree based on five concatenated diploid sequences. B: Maximum-Likelihood tree constructed from Met-II haplotypes.

A: A Neighbour-Joining tree was constructed using concatenated nuclear sequences for all clones Colombian and reference clones. Bootstrap values were calculated as the mean across 1000 randomisations and those >70% are shown for relevant nodes. TcII and TcIII form robust monophyletic clades and all TcV strains are clustered together with the exception of Para6 cl4. Colombian hybrids cluster basally to TcII and TcV with poorly-supported internal structuring. These branches were identified as statistically-incongruent between MLST loci (BIONJ-ILD p<0.001).

B: Maximum-Likelihood topologies were constructed for each locus individually. The phylogeny generated for Met-II, the most polymorphic target, is given as an example. The most appropriate nucleotide substitution model was TrNef+G (three substitution rate categories) based on the AIC. Statistical support for major clades is given as equivalent bootstraps and posterior probabilities from consensus Maximum-Likelihood (1000 pseudo-replicates) and Bayesian trees (based on the HKY+G model), respectively. Haplotypes cluster robustly into two major clades corresponding to TcII and TcIII. Colombian haplotypes were either indistinguishable from or closely-related to those identified in reference TcVI strains, with minimal affinity to either TcII or TcIII Colombia-specific alleles.

For both figures, branch colours indicate isolate DTU (TcII, TcIII, TcV or TcVI). Colombian strain labels are underlined.
**Figure 3.** Unrooted Neighbouring-Joining $D_{AS}$ tree based on 26 microsatellite loci.

$D_{AS}$-based bootstrap values were calculated as the mean across 1000 random diploid re-samplings of the dataset and those >70% are shown for relevant nodes. Branch colours indicate isolate DTU (TcII, TcIII, TcV or TcVI). Colombian strain labels are underlined.

MLMT affords the highest marker resolution; all isolates are characterized by unique MLGs, with the exception of EB cl4 and cl6. All four DTUs are strongly separated in the $D_{AS}$ tree. Within the TcVI clade, all Colombian clones were unique but interspersed among reference TcVI strains from the Southern Cone.
Figure 4. Maximum-Likelihood tree constructed from concatenated maxicircle sequences for 22 Colombian clones and 35 additional isolates from across the Americas.

A Maximum-Likelihood topology was constructed from concatenated maxicircle sequences for all Colombian and reference clones. The most appropriate nucleotide substitution sequences model was GTR+G (nine substitution rate categories) based on the AIC. Statistical support for major clades are given as equivalent bootstraps and posterior probabilities from consensus Maximum-Likelihood (1000 pseudo-replicates) and Bayesian trees (based on the HKY+G model), respectively. Branch colours indicate isolate DTU (TcII, TcIII, TcV or TcVI). Colombian strain labels are underlined.

Isolates group into two monophyletic clades corresponding to TcII and TcIII+TcV+TcVI. Within the latter, all TcVI strains group together, however Colombian hybrid maxicircle haplotypes are diverse between study sites and distinct from reference TcVI isolates. Note strain AACf2 c11 is phylogenetically incongruent between nuclear DAS and ML mitochondrial topologies, appearing to possess a TcV-type maxicircle haplotype.
**Table 1.** Panel of Colombian biological clones and reference clones assembled for analysis.

<table>
<thead>
<tr>
<th>Strain Code</th>
<th>Host/Vector</th>
<th>State</th>
<th>Country</th>
<th>DTU</th>
</tr>
</thead>
<tbody>
<tr>
<td>Colombian Clones</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>EB cl4</td>
<td><em>Homo sapiens</em> neonate (suspected congenital infection)</td>
<td>Boyaca</td>
<td>Colombia</td>
<td>TcII</td>
</tr>
<tr>
<td>EB cl6</td>
<td><em>Homo sapiens</em> neonate (suspected congenital infection)</td>
<td>Boyaca</td>
<td>Colombia</td>
<td>TcII</td>
</tr>
<tr>
<td>EB cl20</td>
<td><em>Homo sapiens</em> neonate (suspected congenital infection)</td>
<td>Boyaca</td>
<td>Colombia</td>
<td>TcII</td>
</tr>
<tr>
<td>PGPA2 cl6</td>
<td><em>Panstrongylus geniculatus</em></td>
<td>Casanare</td>
<td>Colombia</td>
<td>TcII</td>
</tr>
<tr>
<td>PGPA2 cl7</td>
<td><em>Panstrongylus geniculatus</em></td>
<td>Casanare</td>
<td>Colombia</td>
<td>TcII</td>
</tr>
<tr>
<td>PGPA2 cl10</td>
<td><em>Panstrongylus geniculatus</em></td>
<td>Casanare</td>
<td>Colombia</td>
<td>TcII</td>
</tr>
<tr>
<td>CM17</td>
<td><em>Dasypus sp.</em></td>
<td>Carimagua</td>
<td>Colombia</td>
<td>TcIII</td>
</tr>
<tr>
<td>CM25 cl2</td>
<td><em>Dasypus fugilinosa</em></td>
<td>Carimagua</td>
<td>Colombia</td>
<td>TcIII</td>
</tr>
<tr>
<td>SLDN1 cl6</td>
<td><em>Dasypus novemcinctus</em></td>
<td>Casanare</td>
<td>Colombia</td>
<td>TcIII</td>
</tr>
<tr>
<td>TV cl9</td>
<td><em>Triatoma venosa</em></td>
<td>Boyaca</td>
<td>Colombia</td>
<td>TcIII</td>
</tr>
<tr>
<td>AACF2 cl11</td>
<td><em>Canis familiaris</em></td>
<td>Casanare</td>
<td>Colombia</td>
<td>TcVI</td>
</tr>
<tr>
<td>DA cl1</td>
<td><em>Homo sapiens</em></td>
<td>Boyaca</td>
<td>Colombia</td>
<td>TcVI</td>
</tr>
<tr>
<td>DA cl2</td>
<td><em>Homo sapiens</em></td>
<td>Boyaca</td>
<td>Colombia</td>
<td>TcVI</td>
</tr>
<tr>
<td>PG98 cl1</td>
<td><em>Panstrongylus geniculatus</em></td>
<td>Antioquia</td>
<td>Colombia</td>
<td>TcVI</td>
</tr>
<tr>
<td>PG98 cl7</td>
<td><em>Panstrongylus</em></td>
<td>Antioquia</td>
<td>Colombia</td>
<td>TcVI</td>
</tr>
<tr>
<td>Reference Clones</td>
<td>geniculatus</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>------------------</td>
<td>-------------</td>
<td>-----------------</td>
<td>-----------------</td>
<td>-----------------</td>
</tr>
<tr>
<td>Rp540 cl4</td>
<td><em>Rhodnius prolixus</em></td>
<td>Casanare</td>
<td>Colombia</td>
<td>TcVI</td>
</tr>
<tr>
<td>Rp540 cl6</td>
<td><em>Rhodnius prolixus</em></td>
<td>Casanare</td>
<td>Colombia</td>
<td>TcVI</td>
</tr>
<tr>
<td>Rp540 cl7</td>
<td><em>Rhodnius prolixus</em></td>
<td>Casanare</td>
<td>Colombia</td>
<td>TcVI</td>
</tr>
<tr>
<td>Rp540 cl8</td>
<td><em>Rhodnius prolixus</em></td>
<td>Casanare</td>
<td>Colombia</td>
<td>TcVI</td>
</tr>
<tr>
<td>Rp540 cl9</td>
<td><em>Rhodnius prolixus</em></td>
<td>Casanare</td>
<td>Colombia</td>
<td>TcVI</td>
</tr>
<tr>
<td>VS cl6</td>
<td><em>Homo sapiens</em> adult (suspected congenital transmitter)</td>
<td>Boyaca</td>
<td>Colombia</td>
<td>TcVI</td>
</tr>
<tr>
<td>VS cl7</td>
<td><em>Homo sapiens</em> adult (suspected congenital transmitter)</td>
<td>Boyaca</td>
<td>Colombia</td>
<td>TcVI</td>
</tr>
<tr>
<td>VS cl8</td>
<td><em>Homo sapiens</em> adult (suspected congenital transmitter)</td>
<td>Boyaca</td>
<td>Colombia</td>
<td>TcVI</td>
</tr>
<tr>
<td>VS cl10</td>
<td><em>Homo sapiens</em> adult (suspected congenital transmitter)</td>
<td>Boyaca</td>
<td>Colombia</td>
<td>TcVI</td>
</tr>
<tr>
<td>CBB cl2</td>
<td><em>Homo sapiens</em></td>
<td>Tulahuén</td>
<td>Chile</td>
<td>TcII</td>
</tr>
<tr>
<td>Chaco23 col4</td>
<td><em>Triatoma infestans</em></td>
<td>Pr. Hayes</td>
<td>Paraguay</td>
<td>TcII</td>
</tr>
<tr>
<td>Esm cl3</td>
<td><em>Homo sapiens</em></td>
<td>São Felipe</td>
<td>Brazil</td>
<td>TcII</td>
</tr>
<tr>
<td>IVV cl4</td>
<td><em>Homo sapiens</em></td>
<td>Cuncumen</td>
<td>Chile</td>
<td>TcII</td>
</tr>
<tr>
<td>Pot7a cl1</td>
<td><em>Triatoma infestans</em></td>
<td>San Martin</td>
<td>Paraguay</td>
<td>TcII</td>
</tr>
<tr>
<td>Pot7b cl5</td>
<td><em>Triatoma infestans</em></td>
<td>San Martin</td>
<td>Paraguay</td>
<td>TcII</td>
</tr>
<tr>
<td>Sample</td>
<td>Species</td>
<td>Location</td>
<td>Country</td>
<td>Stage</td>
</tr>
<tr>
<td>--------</td>
<td>---------</td>
<td>--------------</td>
<td>----------</td>
<td>-------</td>
</tr>
<tr>
<td>Rita cl5</td>
<td><em>Homo sapiens</em></td>
<td>São Felipe</td>
<td>Brazil</td>
<td>TcII</td>
</tr>
<tr>
<td>T665 cl1</td>
<td><em>Triatoma infestans</em></td>
<td>Pr. Hayes</td>
<td>Paraguay</td>
<td>TcII</td>
</tr>
<tr>
<td>Tu18 cl2</td>
<td><em>Triatoma infestans</em></td>
<td>Tupiza</td>
<td>Bolivia</td>
<td>TcII</td>
</tr>
<tr>
<td>85/847 cl2</td>
<td><em>Dasypus novemcinctus</em></td>
<td>Alto Beni</td>
<td>Bolivia</td>
<td>TcII</td>
</tr>
<tr>
<td>ARMA13 cl1</td>
<td><em>Dasypus novemcinctus</em></td>
<td>Campo Lorro</td>
<td>Paraguay</td>
<td>TcIII</td>
</tr>
<tr>
<td>ARMA18 cl3</td>
<td><em>Dasypus novemcinctus</em></td>
<td>Campo Lorro</td>
<td>Paraguay</td>
<td>TcIII</td>
</tr>
<tr>
<td>JA2 cl2</td>
<td><em>Monodelphis sp.</em></td>
<td>Amazonas</td>
<td>Brazil</td>
<td>TcIII</td>
</tr>
<tr>
<td>M5631 cl5</td>
<td><em>Dasypus novemcinctus</em></td>
<td>Marajo</td>
<td>Brazil</td>
<td>TcIII</td>
</tr>
<tr>
<td>M6421 cl6</td>
<td><em>Homo sapiens</em></td>
<td>Belém</td>
<td>Brazil</td>
<td>TcIII</td>
</tr>
<tr>
<td>SABP19 cl1</td>
<td><em>Triatoma infestans</em></td>
<td>Vitor</td>
<td>Peru</td>
<td>TcIII</td>
</tr>
<tr>
<td>X109/2</td>
<td><em>Canis familiaris</em></td>
<td>Makthlawaiya</td>
<td>Paraguay</td>
<td>TcIII</td>
</tr>
<tr>
<td>X9/3</td>
<td><em>Canis familiaris</em></td>
<td>Makthlawaiya</td>
<td>Paraguay</td>
<td>TcIII</td>
</tr>
<tr>
<td>92.80 cl2</td>
<td><em>Homo sapiens</em></td>
<td>Santa Cruz</td>
<td>Bolivia</td>
<td>TcV</td>
</tr>
<tr>
<td>Bug 2148 cl1</td>
<td><em>Triatoma infestans</em></td>
<td>Rio Grande do Sul</td>
<td>Brazil</td>
<td>TcV</td>
</tr>
<tr>
<td>Chaco2 cl3</td>
<td><em>Triatoma infestans</em></td>
<td>Chaco</td>
<td>Paraguay</td>
<td>TcV</td>
</tr>
<tr>
<td>PAH179 cl5</td>
<td><em>Homo sapiens</em></td>
<td>Chaco</td>
<td>Argentina</td>
<td>TcV</td>
</tr>
<tr>
<td>Para4 cl3</td>
<td><em>Triatoma infestans</em></td>
<td>Paraguari</td>
<td>Paraguay</td>
<td>TcV</td>
</tr>
<tr>
<td>Para6 cl4</td>
<td><em>Triatoma infestans</em></td>
<td>Paraguari</td>
<td>Paraguay</td>
<td>TcV</td>
</tr>
<tr>
<td>Sc43 cl1</td>
<td><em>Triatoma infestans</em></td>
<td>Santa Cruz</td>
<td>Bolivia</td>
<td>TcV</td>
</tr>
<tr>
<td>Vinch101 cl1</td>
<td><em>Triatoma infestans</em></td>
<td>Limari</td>
<td>Chile</td>
<td>TcV</td>
</tr>
<tr>
<td>Chaco17</td>
<td><em>Triatoma infestans</em></td>
<td>Chaco</td>
<td>Paraguay</td>
<td>TcVI</td>
</tr>
<tr>
<td>col1</td>
<td>Species</td>
<td>Location</td>
<td>Country</td>
<td>Virus Type</td>
</tr>
<tr>
<td>-----------</td>
<td>------------------</td>
<td>-------------------</td>
<td>---------</td>
<td>------------</td>
</tr>
<tr>
<td>CL Brener</td>
<td><em>Triatoma infestans</em></td>
<td>Rio Grande do Sul</td>
<td>Brazil</td>
<td>TcVI</td>
</tr>
<tr>
<td>EPV20-1</td>
<td><em>Triatoma infestans</em></td>
<td>Chaco</td>
<td>Argentina</td>
<td>TcVI</td>
</tr>
<tr>
<td>LHVA cl4</td>
<td><em>Triatoma infestans</em></td>
<td>Chaco</td>
<td>Argentina</td>
<td>TcVI</td>
</tr>
<tr>
<td>P251 cl7</td>
<td><em>Homo sapiens</em></td>
<td>Cochabamba</td>
<td>Bolivia</td>
<td>TcVI</td>
</tr>
<tr>
<td>Tula cl2</td>
<td><em>Homo sapiens</em></td>
<td>Tulahuén</td>
<td>Chile</td>
<td>TcVI</td>
</tr>
<tr>
<td>VFRA1 cl1</td>
<td><em>Triatoma infestans</em></td>
<td>Francia</td>
<td>Chile</td>
<td>TcVI</td>
</tr>
</tbody>
</table>
Based on 5 concatenated loci.

| No. of isolates | DP | TE | ST | VS | DP | TE | ST | VS | DP | TE | ST | VS | DP | TE | ST | VS | DP | TE | ST | VS | DP | TE | ST | VS | DP | TE | ST | VS | DP | TE | ST | VS | DP | TE | ST | VS | DP | TE | ST | VS | DP | TE | ST | VS | DP | TE | ST | VS | DP | TE | ST | VS | DP | TE | ST | VS | DP | TE | ST | VS | DP | TE | ST | VS | DP | TE | ST | VS | DP | TE | ST | VS | DP | TE | ST | VS | DP | TE | ST | VS | DP | TE | ST | VS | DP | TE | ST | VS | DP | TE | ST | VS | DP | TE | ST | VS | DP | TE | ST | VS | DP | TE | ST | VS |
|-----------------|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|
Based on 10 concatenated loci

Excludes indels
Table 3. Population genetic parameters calculated from MLMT for different T. cruzi DTUs.

<table>
<thead>
<tr>
<th>DTU</th>
<th>N</th>
<th>G</th>
<th>D_AS ± SD</th>
<th>PL</th>
<th>PA/L ± SE</th>
<th>(\hat{A}_r) ± SE</th>
<th>(\hat{H}_o)</th>
<th>(\hat{H}_e)</th>
<th>%HE</th>
<th>%HD</th>
<th>(G/N)</th>
</tr>
</thead>
<tbody>
<tr>
<td>TcII</td>
<td>14</td>
<td>15</td>
<td>0.44 ± 0.23</td>
<td>24</td>
<td>1.76 ± 0.20</td>
<td>0.58 ± 0.29</td>
<td>0.58</td>
<td>0.65</td>
<td>0.49</td>
<td>0.49</td>
<td>14/15</td>
</tr>
<tr>
<td>TcIII</td>
<td>13</td>
<td>13</td>
<td>0.4 ± 0.15</td>
<td>22</td>
<td>2.35 ± 0.48</td>
<td>0.45 ± 0.12</td>
<td>0.45</td>
<td>0.70</td>
<td>0.49</td>
<td>0.49</td>
<td>13/13</td>
</tr>
<tr>
<td>TcV</td>
<td>8</td>
<td>8</td>
<td>0.15 ± 0.092</td>
<td>22</td>
<td>2.3 ± 0.20</td>
<td>0.85</td>
<td>0.58</td>
<td>54.6</td>
<td>4.5</td>
<td>27.3</td>
<td>8/8</td>
</tr>
<tr>
<td>TcVI</td>
<td>21</td>
<td>21</td>
<td>0.24 ± 0.87</td>
<td>21</td>
<td>0.43 ± 0.12</td>
<td>0.60</td>
<td>0.71</td>
<td>0.49</td>
<td>0.54</td>
<td>41.7</td>
<td>21/21</td>
</tr>
</tbody>
</table>

*Numbers in brackets are indices calculated separately for Colombian clones.*
5.4 Discussion
A clear understanding of the frequency and impact of genetic exchange on the ecological and geographical distributions of *T. cruzi* populations is crucial to establish the epidemiological risk associated with recombinant genotypes and to reconcile the implications hybridization has for *T. cruzi* at both the generational and evolutionary scales. Detecting recombination among natural populations is principally complicated by sampling strategies and marker resolution, considering that strains most likely to be recombining may be highly related and potentially indistinguishable.

The papers included in this chapter describe the characterization of genetic exchange events and hybrid strains from two different parasite populations in North Colombia. Within TcI disease foci, pervasive, indiscriminant mitochondrial introgression was observed affecting ~20% of the study cohort (Ramírez *et al.*, 2012). This is the highest frequency of recombination reported from any natural *T. cruzi* population to date and is probably attributable to the examination of multiple, intensely sampled and minimally subdivided, biological clones per host (Prugnolle and De Meeus, 2010). Similar strategies have now been adopted to explore the sylvatic determinants of TcI genetic diversification in Bolivia (Messenger *et al.*, accepted). The precedent in experimental design established by these studies may represent the most promising intermediary in *T. cruzi* population genetics until imminently superseded by comparative population genomics.

As discussed previously, the failure to detect evidence of reciprocal nuclear hybridization among recombinant Colombian strains may be indicative of an asymmetric, cryptic genetic exchange mechanism, or perhaps more likely, a reflection of the minor amount of nuclear genetic information sampled; without whole nuclear genome sequences for introgression hybrids and putative parental strains, it is impossible to distinguish between these two hypotheses. However, by analogy to other medically-important trypanosome species, the presence of alternate, covert sexual mechanisms within the same species is not entirely unexpected (Rougeron *et al.*, 2009; 2011; Duffy *et al.*, 2013; Hickman *et al.*, 2013; Rogers *et al.*, 2014; Ramírez and Llewellyn, 2014).

By comparison, newly isolated TcVI clones from Colombia satisfied all the expectations of canonical Mendelian F1 progeny, namely, high heterozygosity, minimal genetic diversity, few private alleles and intact parental haplotypes at an approximate 1:1 ratio (Messenger *et al.*, submitted (b)). While not the product of a novel recombination event between local TcII and TcIII strains, the origin(s) of the Colombian hybrids is not unequivocal nor is their status as predecessors to or descendants from Southern Cone TcVI strains. The sharing of mitochondrial haplotypes between southern TcV isolates and a single Colombian TcVI may be the first report of mitochondrial introgression occurring between hybrids lineages. A more parsimonious explanation, based on private alleles and allelic richness measurements, might be that Colombian TcVI strains are siblings which diverged and dispersed northwards, following a single hybridization event between heterozygous TcII and TcIII isolates in the Southern Cone. All six *T. cruzi* DTUs have now been reported from North Colombia at
varying prevalences (Guhl and Ramírez, 2013), incriminating this region as a potential location for prospective hybridization events.

Interestingly, both studies in this chapter identified biparental mitochondrial inheritance as a putative consequence of genetic exchange events. A mosaic maxicircle sequence was detected in a human Colombian TcI isolate and the presence of a recombination breakpoint confirmed by allele-specific PCR. Such a sequence is expected to arise following inter-molecular maxicircle recombination, which necessitates the inheritance of mixed mitochondrial complements. Likewise, sharing of TcV and TcVI mitochondria among Colombian TcVI isolates can be explained if biparental maxicircle inheritance preceded incomplete lineage assortment. Uniparental inheritance of highly heteroplasmic maxicircles might present an indistinguishable scenario but reported mitochondrial heteroplasmy is T. cruzi is thus far low (Messenger et al., 2012) and, at least in the case of the latter isolate, unlikely to bear such homology to a different DTU, even with significant convergent evolution. Parallel observations from experimental crosses of T. b. brucei (Turner et al., 1995; Gibson et al., 2008), suggest that biparental mitochondrial inheritance might be a fundamental, as yet, uncharacterized, biological phenomenon in trypanosomatids.

Importantly, the effects of genetic exchange on parasite phenotype are unknown. The successful establishment of TcV and TcVI among domestic transmission cycles in the Southern Cone, suggests these isolates may possess a phenotypic advantage (heterosis) and/or the ability to outcompete other genotypes. Similarly, if mitochondrial introgression is exploitable as a mechanism of host range extension, recombinants might be expected to present higher mammalian infectivity and growth rates, especially in vectors. The pathological implications of low diversity genotypes in human infections with regards to virulence, transmissibility and drug susceptibility are also largely unknown but warrant consideration, in conjunction with improved methods of detection and isolation of natural hybrids.

The two studies presented in this chapter indicate that natural genetic exchange in T. cruzi is both contemporary and historical, responsible for shaping current parasite population structures, as well as the evolution of distinct T. cruzi DTUs. Together these observations challenge the traditional paradigm of PCE in T. cruzi and highlight the need for additional, intensive and appropriately sampled field surveys, complemented by high resolution, combined nuclear and mitochondrial population genetics analyses.
6. Summary of outputs and future research priorities for Chagas disease

Elucidating the complex epidemiology, phylogeography and clinical variation underlying Chagas disease requires a clear understanding of *T. cruzi* genetic diversity, mechanisms of genetic re-assortment and their impact on natural parasite population structures. The papers herein describe recent developments in *T. cruzi* genotyping techniques and advances in our understanding of parasite ecology and natural genetic exchange mechanisms. The key outputs are:

- Following the establishment of a standardized nuclear MLST protocol, this four marker scheme (*GPI, HMCOAR, RHO1* and *TcMPX*) represents a viable, highly discriminatory and reproducible technique to characterize isolated *T. cruzi* strains to DTU-level; expansion of this panel to include three additional targets (*LAP1, RB19* and *SODB*) has the potential to facilitate intra-lineage discrimination.

- Illumina sequencing of the maxicircle genome from the TcI reference strain (Sylvio X10/1) at 183X coverage, provided the first evidence of heteroplasmy in the mitochondrial coding region at a ~10 fold lower abundance compared to the consensus genome. It is likely that the frequency of minor heteroplasmic maxicircles may be higher in recently sampled field strains, which have been minimally sub-passaged. However, this level of intra-strain maxicircle diversity does not appear to be sufficiently divergent to adversely affect genotyping using multi-copy mitochondrial genes.

- Combined mitochondrial MLST and MLMT affords the highest intra-lineage resolution with the advantages of identifying cryptic genetic exchange events, undetectable by nuclear loci, as evidenced by nuclear-mitochondrial phylogenetic incongruence. Among geographically dispersed TcI populations, multiple intra-TcI and inter-DTU (TcI and TcIII/TcIV) mitochondrial introgression events were observed without any apparent reciprocal nuclear hybridization.

- Illumina amplicon deep sequencing of clinical isolates revealed an unprecedented level of intra-host parasite multiclonality and highlighted potential diversifying selection affecting antigenic surface proteases; genetic diversity within this multi-gene family may facilitate survival in the mammalian host. However, this study was unable to identify any relationship between multiplicity of infection and patient sex, age or clinical manifestations.

- High resolution nuclear and mitochondrial genotyping of contemporaneous TcI clones from Bolivia demonstrated that ecological host fitting was the predominant mechanism structuring parasite populations among arboreal and terrestrial transmission cycles. Reduced genetic diversity, accelerated parasite dissemination between densely populated areas and mitochondrial gene flow between domestic and sylvatic parasite populations,
suggests humans may have played a crucial role in *T. cruzi* dispersal across the Bolivian highlands.

- The significant impact human activity can have on parasite population structures was also observed in Brazil where reduced TcI genetic diversity among Atlantic Forest populations was attributed to ongoing anthropogenic habitat fragmentation.

- Genetic characterization of Tc-IDOM isolates in comparison to TcI strains from across its endemic range, indicated that these human-associated, genetically homogeneous genotypes likely originated in North/Central America before dissemination southwards. Molecular dating of Tc-IDOM clones from Colombia confirmed that this clade emerged 23,000 ± 12,000 years, coinciding with the earliest human migration into South America.

- Gross nuclear-mitochondrial phylogenetic incongruence affecting ~20% of TcI clones from Colombia suggests that mitochondrial introgression is a common mechanism of genetic exchange among natural parasite populations. Similar intra-TcI introgression events were also observed at lower frequencies among intensively-sampled sylvatic transmission cycles in Bolivia and Brazil.

- A mosaic maxicircle detected in a Colombia patient may be the result of inter-molecular mitochondrial recombination, suggesting biparental mitochondrial inheritance can occur during some genetic exchange events. This phenomenon would also explain the mitochondrial heteroplasmy observed in the TcI mitochondrial genome sequence, as following hybridization, one maxicircle population is lost stochastically during mitotic division, but may never be eliminated entirely.

- Hybrid isolates in Colombia were distinct from, but related to, TcVI reference strains from the Southern Cone, incriminating them as *'bona fide'* TcVI clones and confirming that the geographical range of this DTU is much more extensive than previously assumed. The origin(s) of these Colombian hybrids remains unresolved; the sharing of mitochondrial haplotypes between southern TcV strains and a Colombian TcVI clone, may be explained if Colombian hybrids are a sibling group to southern TcVI isolates, which diverged and dispersed northwards, following a single hybridization event with biparental mitochondrial inheritance between heterozygous TcII and TcIII isolates.

The results described herein highlight a number of prospective avenues of investigation. Establishing any relationship between *T. cruzi* genetic diversity and clinical outcome will require significant improvements in both clinical genotyping and patient sampling. Detection of multiple distinct parasite clones within Colombian patients by combined mtMLST and MLMT, indicated that super-infection may be a common feature of some transmission foci. It is highly probable that those patients sampled as part of the Illumina deep sequencing study were not recently exposed to high forces of infection given vector transmission had been interrupted in Goiás, Brazil 10-20 years ago (Marsden *et al.*, 1994) and severely reduced in Cochabamba, Bolivia in the mid-2000s (Espinoza *et al.*, 2014).
This strategy to examine intra-patient multiclonality warrants additional evaluation, preferably within hyperendemic populations, in comparison with age-matched cohorts without vector exposure. If multiplicity of infection is associated with severity of cardiac disease, this might alone, or in combination with other blood-based biomarkers of disease progression (Requena-Méndez et al., 2013), justify targeted benznidazole treatment among the subset of chronic adults at high risk of mortality (provided efficacy to prevent advanced cardiomyopathy in adults is demonstrated by the BENEFIT and TRAENA trials (Marin-Neto et al., 2009)).

Furthermore, application of this technology to investigate the interaction between parasite multiclonality and host immune response among longitudinal groups of patients could enhance our understanding T. cruzi immune evasion. Unlike African trypanosomiasis, where mechanisms of antigenic variation of variant surface glycoproteins (VSGs) are well established (Borst and Cross, 1982; Pays, 2005, Morrison et al., 2009b; Jackson et al., 2012), the role of the large, highly repetitive, surface molecule gene families, which make up ~50% of the T. cruzi genome (Andersson, 2011), in parasite persistence, is largely unknown. With improved genome assemblies of reference strains (Weatherly et al., 2009; Aslett et al., 2010), it is now feasible to reconstruct the more conserved gene families into chromosomal contigs that could be used to examine genomic architectural re-arrangements and/or variations in copy number of antigenic surface molecules during chronic infections, when combined with sequential patient sampling (Urban et al., 2011; Minning et al., 2011; Sterkers et al., 2011; Pavia et al., 2012).

All of these potential studies are contingent on parallel improvements in direct clinical genotyping. As demonstrated by the Illumina study, PCR amplification and parasite hemoculturing steps are not ideal and introduced an unquantifiable but unavoidable number of biases. The exploitation of recent techniques designed to enrich pathogen DNA for field-based genome sequencing of other parasite protozoa, such as selective methylation-dependent degradation of human DNA (Oyola et al., 2013), whole genome bead-capture (Bright et al., 2012), RNA aptamer-based ligand capture (Nagarkatti et al., 2012) and WGA, (Morrison et al., 2007; Nair et al., 2014) should be considered for future T. cruzi clinical sampling.

To improve our knowledge of sylvatic transmission dynamics, mammalian host associations and natural parasite population structures, landscape genetics (Biek and Real, 2010; Manel and Holderegger, 2013), combined with intensive spatio-temporal sampling, may represent a promising new strategy, especially if complemented with parasite whole genome sequencing (WGS), as has been initiated with field isolates of Leishmania (Downing et al., 2011; Rogers et al., 2014). All study sites described herein merit further sampling efforts, particularly the ecologically-rich biomes of Brazil and Bolivia and domestic transmission cycles in North Colombia.

The mechanisms of genetic exchange identified among field populations are challenging to reconcile with those observed experimentally (Ramirez and Llewellyn, 2014). Among the majority of field data presented herein, mitochondrial introgression was detected as a
common feature of natural transmission cycles (Messenger et al., 2012; Ramírez et al., 2012; Zumaya-Estrada et al., 2012; Lima et al., 2014; Messenger et al., accepted); the failure to detect any reciprocal nuclear hybridization likely reflects the amount of each genome sampled (20% of the mitochondrial vs. <0.1% of the nuclear genome). To determine whether a genuine alternate mating system, involving asymmetric mitochondrial inheritance, exists in T. cruzi, WGS of introgression isolates in comparison with putative nuclear parents and mitochondrial donors could be undertaken. Resulting genome-wide SNP and heterozygosity distributions, copy number variations and chromosomal arrangements could be used to resolve the relative contribution of nuclear material to hybrid strains and to characterize the patterns of genetic inheritance following recombination.

To investigate whether meiosis is actively occurring as a mechanism of genetic exchange, considering the resemblance of hybrid DTUs to Mendelian F1 progeny, a similar approach to Peacock et al., 2011; 2014 could be attempted in T. cruzi, involving fluorescent tagging of the N-termini of conserved meiotic orthologues known to function during recombination in T. b. brucei (DMC1, HOP1, MND1 and SPO11). Tracking of meiosis gene expression throughout in vitro and in vivo life cycles may resolve their contributions to genetic exchange mechanisms or potentially expose novel protein functions; C. albicans exploits SPO11 to facilitate recombination between sister chromatids during its non-meiotic parasexual cycle (Forche et al., 2008) and Giardia co-opts its meiotic orthologues to undergo homologous recombination during karyogamy (Poxleitner et al., 2008).

Importantly, the effects of genetic exchange on parasite phenotype are unknown. Few studies have focused on phenotyping recently-isolated T. cruzi field strains; most are reliant on historical reference collections, whose behaviour is known to change after significant time in culture (Moreira et al., 2012), and those performed on handfuls of newer isolates have reported significant variation between DTUs and even among closely related strains (Andrade and Magalhães, 1997; Martínez-Díaz et al., 2001).

In vitro characterization of both parental and hybrid T. cruzi strains has the potential to reveal inheritance patterns of phenotypic traits and facilitate accurate assessments of the epidemiological risk associated with recombinant genotypes. To develop high-throughput, T. cruzi phenotyping, an analogous approach to those used for genetic crosses of other trypanosomatids could be adopted; parasite clones expressing either a fluorescent (Akopyants et al., 2009; Sadlova et al., 2011; Bouvier et al., 2013; Inbar et al., 2013) or bioluminescent reporter gene (Henriques et al., 2012; 2014; Lewis et al., 2014), could be evaluated for parameters such as in vitro growth rate, mammalian cell infectivity, rate of metacyclogenesis, vector permmissibility among different species of triatomine bug, and in vivo parasitaemia, infection course, tissue sequestration and drug susceptibility. Advantages of using a transgenic reporter system include easier and more accurate parasite quantification by FACs and highly sensitive real-time imaging of infection experiments (Lewis et al., 2014). This technique could be used to compare the biological behaviour of a number of epidemiological important T. cruzi genotypes. For example, hybrid (TcV and TcVI) and parental representatives (TcII and TcIII), TcDOM vs. sympatric sylvatic TcI isolates as well as
strains that have undergone mitochondrial introgression compared to their nuclear parents and mitochondrial donors. This system could then be expanded to investigate the interaction between distinct genotypes, expressing different reporter genes and selectable markers, during competitive mixed or sequential infections, in vitro, in vectors and in mice, with the potential to recover novel hybrid strains under double drug pressure.

Lastly, unraveling mitochondrial inheritance is complicated due to our inability to generate a transgenic cell line with all multi-copy maxicircles fluorescently tagged. An alternate approach to directly address the cytological mechanisms of mitochondrial segregation to progeny cells, might be to isolate individual parasites by FACs, immediately following genetic crosses, amplify genetic material by WGA, and characterize using single cell genomics (as has been recently pioneered to dissect multiclonal malaria infections (Nair et al., 2014)).

This PhD project contributed to the further development of new methods to resolve the complex population genetics of *T. cruzi*, specifically, the establishment of the first panel of mitochondrial MLST markers, formalization of a transferable nuclear MLST scheme and design of Illumina deep sequencing markers. Application of the latter to patient isolates revealed extensive *T. cruzi* intra-host multiclonality, with implications for enhanced understanding of transmissibility and clinical presentation. Similarly, analyses of field populations using a combination of mitochondrial and nuclear MLST and MLMT, provided new insights into the distribution of major genotypes and their associations with diverse domestic transmission cycles and sylvatic ecologies. Unanticipated occurrence of widespread intra- and inter-DTU mitochondrial introgression was observed, reaffirming the importance of genetic exchange in influencing the evolution of *T. cruzi*.

There is intense international research interest in trypanosomatids because of their shared unusual, and often, unique biological features, as well as considerable public health importance. The outputs from this PhD have the potential to contribute to resolving several enigmas of *T. cruzi* biology, including, a comprehensive understanding of parasite phylogeography; the relationship between *T. cruzi* DTUs, parasite multiclonality and congenital transmission or clinical prognosis; the identification of novel, non-canonical genetic mechanisms; and the epidemiological implications of parasite hybridization.

Some of the emerging findings from *T. cruzi* research are reminiscent of observations from other pathogens, for example the parasexual and sexual genetic exchange mechanisms of fungi, and the epidemiology of clonal and hybrid lineages of *Toxoplasma gondii*. It is important to sustain and nurture dialogue between research disciplines, across pathogens, and through international networks and technology transfer. The ability to improve our understanding of the genetic diversity of *T. cruzi* and other pathogens is clearly demonstrated by outputs of this research project. In addition to the prospective studies described herein, the future research agenda for Chagas disease needs to be considered in the context of changing disease epidemiology. Over the last 20-30 years, concomitant with successful trans-national vector control programmes and significant infrastructure and social development throughout
Latin America, Chagas disease has shifted from a neglected endemic disease of the rural poor to an urbanized chronic infection and now a potentially emergent global health problem.

It is estimated that by 2030, 90% of the population in Latin American will inhabit cities (Kates and Parris, 2003), where chronic indeterminate individuals, infected decades ago, are of particular concern. In the absence of affordable, efficacious chemotherapy for chronic disease, research priorities must focus on the development of methods to detect the 20-30% of patients at risk of disease progression and the evaluation of novel compounds for chronic stage treatment. In parallel, congenital transmission among such populations now accounts for 25% of new infections (Jannin and Salvatella, 2006), is challenging to preempt, but comparatively easy to control if identified early.

The rapid advancements in molecular biology technology, genomics, proteomics, microbiomics, etc, and parallel plummeting reagent costs, present the ideal opportunity to adopt a multidisciplinary approach to expand our academic understanding of *T. cruzi* and translate these improvements for the direct benefit of Chagas disease patients. Until then, presented herein are some of the highest resolution genotyping techniques developed in *T. cruzi* to date which have the potential to expand our current understanding of parasite genetic diversity and its relevance to natural population structures and the clinical outcome of Chagas disease.


**infestans** (Reduviidae, Triatominae) in the valleys of La Paz, Bolivia. *The American Journal of Tropical Medicine and Hygiene*, 82, 574-579.


geniculatus (Hemiptera: Reduviidae) in Caracas (Metropolitan District) and neighbouring states, Venezuela. International Journal for Parasitology, 35, 1379-1384.


Da Silveira Pinto, A., de Lana, M., Britto, C., Bastrenta, B. & Tibayrenc, M. (2000). Experimental *Trypanosoma cruzi* biclonal infection in *Triatoma infestans*: detection of


Franzén, O., Ochaya, S., Sherwood, E., Lewis, M.D., Llewellyn, M.S., Miles, M.A. &


Gomes, Y.M., Lorena, V.M. & Luquetti, A.O. (2009). Diagnosis of Chagas disease: what has been achieved? What remains to be done with regard to diagnosis and follow up studies? Memórias do Instituto Oswaldo Cruz, 104, Suppl 1, 115-121.


Henriksson, J., Porcel, B.M., Rydaker, M., Sabaj, V., Galanti, N., Ruiz, A.M., Cazzulo, J.J.,
linkage groups in spite of extensive chromosomal size variation in *Trypanosoma cruzi.*
*Molecular and Biochemical Parasitology,* 73, 63-74.

Henriksson, J., Dujardin, J.C., Barnabé, C., Brisse, S., Timperman, G., Venegas, J.,
*Trypanosoma cruzi* is mainly progressive and is evolutionarily informative. *Parasitology,*
124, 277-286.

Bioluminescent imaging of *Trypanosoma cruzi* infection in *Rhodnius prolixus.* *Parasites and Vectors,* 5, 214.

(2014). *In vivo* imaging of mice infected with bioluminescent *Trypanosoma cruzi* unveils
novel sites of infection. *Parasites and Vectors,* 7, 89.

Hermann, E., Truyens, C., Alonso-Vega, C., Rodriguez, P., Berthe, A., Torrico, F. & Carlier,
Y. (2004). Congenital transmission of *Trypanosoma cruzi* is associated with maternal
enhanced parasitemia and decreased production of interferon-gamma in response to parasite
antigens. *Journal of Infectious Diseases,* 189, 1274-1281.

Herrera, C., Barques, M.D., Fajardo, A., Montilla, M., Triana, O., Vallejo, G.A. & Guhl, F.
(2007b). Identifying four *Trypanosoma cruzi* I isolate haplotypes from different geographic
regions in Colombia. *Infection, Genetics and Evolution,* 7, 535-539.

Herrera, C., Guhl, F., Falla, A., Fajardo, A., Montilla, M., Adolfo Vallejo, G., & Dolores

Herrera, H.M., Lisboa, C.V., Pinho, A.P., Olifiers, N., Bianchi, R.C., Rocha, F.L., Mourão,
reservoir host for the main lineages of *Trypanosoma cruzi* in the Pantanal region, Brazil.
*Transactions of the Royal Society of Tropical Medicine and Hygiene,* 102, 1133-1139.

played by sympatric collared peccary (*Tayassu tajacu*), white-lipped peccary (*Tayassu pecari*), and feral pig (*Sus scrofa*) as maintenance hosts for *Trypanosoma evansi* and
*Trypanosoma cruzi* in a sylvatic area of Brazil. *Parasitology Research,* 103, 619-624.

(2005). *Trypanosoma cruzi* infection in wild mammals of the National Park ‘Serra da
Capivara’ and its surroundings (Piauí, Brazil), an area endemic for Chagas disease.
*Transactions of the Royal Society of Tropical Medicine and Hygiene,* 99, 379-388.


Higuera, S.L., Gubl, F. & Ramírez, J.D. (2013). Identification of *Trypanosoma cruzi* discrete typing units (DTUs) through the implementation of a high-resolution melting (HRM) genotyping assay. *Parasites and Vectors*, 6, 112.


Keenan, M., Abbott, M.J., Alexander, P.W., Armstrong, T., Best, W.M., Berven, B., Botero,


Noireau, F. (2009b). Wild *Triatoma infestans*, a potential threat that needs to be monitored. *Memórias do Instituto Oswaldo Cruz*, 104, Suppl 1, 60-64.


Pinto, C.M., Kalko, E.K.V., Cottontail, I., Wellinghamusen, N. & Cottontail, V.M. (2012). TcBat a bat-exclusive lineage of *Trypanosoma cruzi* in the Panama Canal Zone, with
comments on its classification and the use of the 18S rRNA gene for lineage identification. Infection, Genetics and Evolution, 12, 1328-1332.


Appendix A
COVER SHEET FOR EACH ‘RESEARCH PAPER’ INCLUDED IN A RESEARCH THESIS

Please be aware that one cover sheet must be completed for each ‘Research Paper’ included in a thesis.

1. For a ‘research paper’ already published

1.1. Where was the work published? ....................................................................................................................

1.2. When was the work published? ....................................................................................................................

1.2.1. If the work was published prior to registration for your research degree, give a brief rationale for its inclusion
......................................................................................................................................................................
......................................................................................................................................................................

1.3. Was the work subject to academic peer review? ............................................................................................

1.4. Have you retained the copyright for the work? Yes / No
If yes, please attach evidence of retention.
If no, or if the work is being included in its published format, please attach evidence of permission from copyright holder (publisher or other author) to include work

2. For a ‘research paper’ prepared for publication but not yet published

2.1. Where is the work intended to be published? Parasite Genomics Protocols. Methods in Molecular Biology

2.2. Please list the paper’s authors in the intended authorship order
Message, L.A., Yeo, M., Lewis, M.D., Llewellyn, M.S. & Miles, M.A.

2.3. Stage of publication – Not yet submitted / Submitted / Undergoing revision from peer reviewers’ comments (In press)

3. For multi-authored work, give full details of your role in the research included in the paper and in the preparation of the paper. (Attach a further sheet if necessary)
The candidate drafted the manuscript in close consultation with the other co-authors.
......................................................................................................................................................................

NAME IN FULL (Block Capitals) LOUISA ALEXANDRA MESSENGER

STUDENT ID NO: 223021

CANDIDATE’S SIGNATURE ___________________________ Date 10/2/15

SUPERVISOR/SENIOR AUTHOR’S SIGNATURE (3 above) ___________________________
Molecular Genotyping of Trypanosoma cruzi for Lineage Assignment and Population Genetics

Louisa A. Messenger, Matthew Yeo, Michael D. Lewis, Martin S. Llewellyn, and Michael A. Miles

Abstract

Trypanosoma cruzi, the etiological agent of Chagas disease, remains a major public health problem in Latin America. Infection with T. cruzi is lifelong and can lead to a spectrum of pathological sequelae ranging from subclinical to lethal cardiac and/or gastrointestinal complications. Isolates of T. cruzi can be assigned to six genetic lineages or discrete typing units (DTUs), which are broadly associated with disparate ecologies, transmission cycles, and geographical distributions. This extensive genetic diversity is also believed to contribute to the clinical variation observed among chagasic patients. Unravelling the population structure of T. cruzi is fundamental to understanding Chagas disease epidemiology, developing control strategies, and resolving the relationship between parasite genotype and clinical prognosis.

To date, no single, widely validated, genetic target allows unequivocal resolution to DTU-level. In this chapter we present standardized methods for strain DTU assignment using PCR-restriction fragment length polymorphism analysis (PCR-RFLP) and nuclear multilocus sequence typing (MLST). PCR-RFLPs have the advantages of simplicity and reproducibility, requiring limited expertise and few laboratory consumables. MLST data are more laborious to generate but more informative; DNA sequences are readily transferrable between research groups and amenable to recombination detection and intra-lineage analyses. We also recommend a mitochondrial (maxicircle) MLST scheme and a panel of 28 microsatellite loci for higher resolution population genetics studies.

Due to the scarcity of T. cruzi in blood and tissue, all of these genotyping techniques have limited sensitivity when applied directly to clinical or biological specimens, particularly when targets are single (MLST) or low copy number (PCR-RFLPs). We therefore describe essential protocols to isolate parasites, derive biological clones, and extract T. cruzi genomic DNA from field and clinical samples.

Key words Trypanosoma cruzi, PCR, Genotyping, Phylogenetics, Microsatellites, MLST, RFLP, Mitochondria, Sequencing

1 Introduction

Chagas disease is the most important parasitic infection in Latin America, where an estimated 10–12 million individuals are infected, with a further 80 million at risk [1]. The etiological agent, Trypanosoma cruzi, is a complex zoonosis, with a broad
endemic range that extends from the southern United States to Argentinean Patagonia. Disease transmission primarily occurs in areas where humans are exposed to the contaminated feces of domiciliated triatomine bug vectors. In the absence of chemotherapy, infection with *T. cruzi* is life-long and can lead to a spectrum of pathological sequelae ranging from subclinical to debilitation and death by irreversible cardiac and/or gastrointestinal syndromes [2]. Diagnosis and treatment options are further complicated by disproportionate distributions of disease pathologies; cardiomyopathies occur throughout South and Central America, whereas gastrointestinal complications are more common south of the Amazon. It has been suggested that this geographical heterogeneity is associated with genetic variation among *T. cruzi* strains [3–5]. However, the relationship between parasite genotype and clinical outcome remains controversial.

*T. cruzi* displays remarkable genetic diversity and a range of markers can be used to delineate this species. Typing of genetic polymorphisms in conserved housekeeping genes can define major genetic lineages [6–8], while analysis of hypervariable loci, such as microsatellites [9–11] or kDNA minicircle sequences [12–14], potentially allows identification of profiles specific to individual strains. Historically, the study of *T. cruzi* has been hindered by a lack of standardized molecular typing methods and the use of various alternative nomenclatures (recently reviewed in [15]). One useful conceptual development has been that of the discrete typing unit (DTU) which groups isolates using shared molecular characteristics but without explicitly defining their evolutionary relatedness [16]. For *T. cruzi* multilocus genotyping has consistently identified six DTUs, which are each correlated with distinct but not exclusive ecologies and geographical distributions [17]. Recently, DTU nomenclature has been revised by international consensus to reflect the current understanding of *T. cruzi* genetic diversity [18].

Molecular analyses suggest that *T. cruzi* has a predominantly clonal population structure, punctuated by infrequent genetic exchange events. DTUs TcI-TcIV form monophyletic clades and TcV and TcVI are known to be recent inter-lineage hybrids [19]. Multilocus sequence typing (MLST) data support these designations with TcI-TcIV characterized by substantial allelic homozgyosity, likely resulting from recurrent, genome-wide and dispersed gene conversion, while TcV and TcVI display natural heterozygosity and minimal distinction, sharing intact alleles from their parental progenitors (TcII and TcIII) [20–22]. The origin(s) of these hybrid lineages is unresolved and it is presently contested whether they arose from two independent genetic exchange events [19, 23], or a single incidence of hybridization followed by clonal divergence [24] (recently reviewed in [25]).

The epidemiological relevance of the *T. cruzi* DTUs has also been the subject of considerable debate, with evidence emerging
to support historical and contemporary associations of particular lineages with different transmission ecologies. In general, TcI, TcII, TcV, and TcVI are frequently isolated from domestic cycles and are responsible for the majority of human infections. The distribution of domestic TcI extends from the Amazon Basin northwards, where it is the primary cause of Chagas disease in Venezuela and Colombia [26, 27]. TcI is also ubiquitous among arboreal sylvatic transmission cycles throughout Latin America [28, 29], and commonly isolated from Didelphis species and the triatomine tribe Rhodniini [30]. By contrast, TcII, TcV, and TcVI appear restricted to domestic transmission in southern parts of South America. Strains from these three DTUs are rarely isolated from sylvatic reservoirs and their ecological niches are largely undefined [17]. TcIII has a dispersed terrestrial distribution that ranges from Amazonia to Argentina, where it is primarily transmitted by Panstrongylus geniculatus to Dasypus novemcinctus and other burrowing mammals [31–33]. TcIV is poorly understood, principally because several genotyping methods fail to distinguish this lineage from others, particularly from TcIII [6]. However, TcIV is known to circulate sympatriically with TcI in wild primates [34] and raccoons [29] in Amazonia and North America, respectively. It is also increasing in epidemiological importance and has been implicated in recent oral outbreaks in Amazonia [34, 35] and as a secondary agent of Chagas disease in Venezuela [3]. As yet, TcIII and TcIV only sporadically invade domestic transmission cycles, but this may reflect inadequate and/or inappropriate sampling and the insensitivity of conventional genotyping methods. Furthermore, some of these ecological associations are complicated by overlapping sylvatic and domestic transmission cycles and frequent mixed infections in individual humans [36, 37], mammalian reservoirs [32, 38], and triatomine vectors [8, 39–41].

Elucidating the population structure and genetic diversity of T. cruzi is critical to furthering our understanding of the complex transmission dynamics, clinical variability and phylogeography underlying Chagas disease. Secondly, detecting recombination among T. cruzi populations is also of profound epidemiological importance considering the expansion of the hybrid lineages within the domestic niche and the capacity for genetic exchange to drive the evolution of novel virulent recombinant strains. As yet, no single marker affords complete, unequivocal DTU resolution, and reliance on only one target is inadvisable given the potential confounding influence of hybridization [12, 21]. In this chapter we describe genotyping methods to assign T. cruzi isolates to DTU-level and those that can be used for higher resolution intra-lineage diversity studies.

For optimal genotyping results we strongly recommend the use of biologically cloned material, wherever possible. Multiclonality within individual T. cruzi strains can manifest as mixed infections
of different DTUs [37–39, 41, 42] or multiple variants of the same genetic lineage [41, 43]. Infra-population genetic diversity is largely determined by levels of super-infection from discrete sources [44], inbreeding among closely related genotypes [45] and simultaneous transmission of multiclonal populations between hosts [38]. We describe routine protocols to isolate *T. cruzi* parasites from infected patients/mammals and triatomine bugs. We then recommend methods to derive biological clones from *T. cruzi* strains, including plating on a solid medium [41], limiting dilution or micromanipulation of individual parasites [46] and also suggest techniques to extract genomic DNA from resulting axenic cultures as well as directly from clinical and field isolates.

To genotype *T. cruzi* isolates to DTU-level we recommend a standardized triple-assay comprising PCR product size polymorphism analysis of the 24Sα rRNA gene (LSU rDNA) and PCR-restriction fragment-length polymorphism analysis (PCR-RFLP) using heat shock protein 60 (*HSP60*) and glucose-6-phosphate isomerase (*GPI*) [47]. These PCR-based assays have the advantages of being easily reproducible and implemented with limited expertise, technical resources, and sample material. However, this methodology was developed using a panel of biologically cloned reference isolates and is reliant on the presence/absence of specific single-nucleotide polymorphisms (SNPs) and may be insensitive to mutations in as yet untested strains. In addition, both PCR-RFLPs are based on low copy targets and were evaluated using culture-extracted DNA and thus their sensitivity against field or clinical specimens and for resolving mixed infections may vary. The repertoire of PCR-based *T. cruzi* genotyping techniques is ever expanding and those recently described by D’Avila et al. [48], Burgos et al. [49], and Van der Auwera et al. [50] may be more appropriate for the aforementioned sample types.

Another technique that we advocate to unambiguously assign isolates to DTU-level is nuclear multilocus sequence typing (MLST). This is a sequence-based approach, which exploits conserved nucleotide diversity present in four single-copy housekeeping genes (3-hydroxy-3-methylglutaryl-CoA reductase (*HMCOAR*), glucose-6-phosphate isomerase (*GPI*), mitochondrial peroxidase (*TcMPX*), and rho-like GTP binding protein (*RH01*)) [20, 51] and can be used as an adjunct to DTU allocation, in the rare cases when PCR-RFLPs fail to unequivocally genotype samples. MLST data offer minimal subjectivity in analysis and are transferable and electronically portable, allowing for inter-laboratory comparisons without the exchange of reference isolates. Our research group, along with others [51], is presently expanding this panel of loci with the aim of formalizing an MLST scheme that can be used for high resolution genetic diversity studies [52].

We anticipate that with the rapid advancement of sequencing technology, current genotyping methods will imminently be
superseded by comparative genomics of multiple representatives from each *T. cruzi* DTU [53]. However, in the interim, we recommend the use of a panel of 28 microsatellite loci (multilocus microsatellite typing, MLMT) and ten mitochondrial gene fragments (maxicircle MLST) to address intra-lineage population genetic hypotheses using appropriately assembled isolate cohorts. Microsatellites are short, neutrally evolving, codominant tandem repeats, with mutation rates several orders of magnitude higher than protein-coding genes [54]. These hypervariable markers provide a method of identifying and tracking individual strains as well as assessing the frequency of alleles in a given population. This MLMT scheme is highly discriminatory and has previously been used to describe intra-TcI and -TcIII population structuring on a continental scale [10, 33], to reveal genetic exchange within TcI domestic/peridomestic populations in Ecuador [11] and to expose the role of mammalian reservoirs in the diversification of *T. cruzi* genotypes [38]. Potential drawbacks associated with MLMT include limited transferability between laboratories and genotyping errors arising from homoplasy (when alleles are identical in sequence but not descent), allelic dropout, misprinting, artifact peaks, and stutter patterns [55]. Maxicircle MLST exploits inherent features of mitochondrial DNA, specifically uniparental inheritance and a faster mutation rate (compared to nuclear DNA), to detect directional gene flow among closely related isolates. Maxicircle MLST can be used in parallel with nuclear loci (MLMT and/or nuclear MLST) to identify phylogenetic incongruence, which is indicative natural recombination. This combined approach has uncovered novel mitochondrial introgression events occurring across geographically dispersed TcI populations [56] and revealed pervasive genetic exchange within Colombian TcI transmission cycles [44].

Herein, we describe the protocols used to (1) isolate *T. cruzi* samples from infected patients, mammalian hosts and triatomine bugs, (2) derive biological clones from *T. cruzi* strains by micro-manipulation, plating on solid medium, or limiting dilution, (3) extract parasite DNA from cultured epimastigotes, human/mammalian hemocultures, or triatomine bug intestinal homogenates, (4) assign isolates to DTU-level using PCR-RFLP analysis, (5) amplify, sequence and analyze nuclear and maxicircle MLST targets, and (6) amplify, multiplex, and analyze microsatellite allele sizes.

## 2 Materials

Prepare all solutions using ultrapure water (purify deionized water to attain a resistivity of 18 MΩ cm at 25 °C) and analytical grade reagents.
All experimental work which involves handling live *T. cruzi* parasites should be conducted in a designated laboratory and in accordance with locally approved standard operating procedures (SOPs). All manipulation of live material should be carried out within a Class II microbiological safety cabinet. Accidental infection with *T. cruzi* can arise from inoculation of a single infectious metacyclic trypomastigote or bloodstream-form trypomastigote and at least sixty-five cases of laboratory transmission have been recorded [57]. However, the risk of laboratory-acquired infection is minimal if appropriate guidelines are adhered to (see Note 1).

### 2.1 Isolation of *T. cruzi*

Here we present possible protocols for the isolation of *T. cruzi*, techniques for biologically cloning resulting parasites and methods of extracting *T. cruzi* genomic DNA. Choice of technique will depend upon the original source of the parasite and quality of DNA template required for downstream applications (see Note 2).

To maximize the likelihood of isolate recovery and minimize loss of clonal diversity, we strongly recommend processing all field and clinical samples by simultaneously (1) inoculating strains into axenic culture (proceed to (2) before the first re-passage), (2) biologically cloning strains, and (3) directly extracting genomic DNA (see Fig. 1).

#### 2.1.1 Direct Hemoculture from Patients/Mammals

1. Blood agar base (Sigma-Aldrich, UK).
2. Agar (Sigma-Aldrich, UK).
3. Tryptone (Sigma-Aldrich, UK).

---

**Fig. 1** Schematic of *T. cruzi* strain isolation, cloning, and DNA extraction protocols
4. Sodium chloride, NaCl (Sigma-Aldrich, UK).
5. Sterile defibrinated rabbit blood.
6. Gentamycin (Sigma-Aldrich, UK).
7. 5-Fluorocytosine (Sigma-Aldrich, UK).
8. Liver infusion broth (Difco™, Becton Dickinson, USA).
9. Glucose (Sigma-Aldrich, UK).
10. Potassium chloride, KCl (Sigma-Aldrich, UK).
11. Disodium hydrogen phosphate, Na₂HPO₄ (Sigma-Aldrich, UK).
12. Hemin (Sigma-Aldrich, UK).
13. Sodium hydroxide, NaOH (Sigma-Aldrich, UK).
15. Ketamine hydrochloride (Sigma-Aldrich, UK).
16. Absolute ethanol (analytical reagent grade).
17. Iodine (Sigma-Aldrich, UK).
18. Guanidine hydrochloride (Sigma-Aldrich, UK).
19. Ethylenediaminetetraacetic acid (EDTA) disodium salt dihydrate (Sigma-Aldrich, UK).
20. Refrigerated centrifuge.
21. Sterile 15 ml centrifuge tubes (Greiner Bio-One, UK).
22. Sterile Nunclon™ ∆ flat sided tubes (#734-2068, Nunc, UK).
23. Rubber caps from sodium heparin vacutainer tubes (#368480, Scientific Laboratory Supplies, UK).
24. Parafilm (VWR, UK).
25. Sterile 1, 2, 5, and 20 ml BD Plastipak™ syringes with needles (Becton Dickinson, USA).
26. BD Vacutainer® plus plastic K₂ EDTA tubes (Becton Dickinson, USA).
27. 28 °C humidified incubator.
28. Inverted microscope.
29. Sterile glycerol (VWR, UK).
30. Sterile cryovials (Nunc, Denmark).

2.1.2 Isolation from Triatomine Bugs (Xenodiagnosis)

1. Uninfected triatomine bug colony.
2. Mercuric chloride, HgCl₂ (Sigma-Aldrich, UK).
3. Hydrochloric acid sp.gr.1.18, HCl (VWR, UK).
4. Sodium chloride, NaCl (Sigma-Aldrich, UK).
5. Absolute ethanol (analytical reagent grade).
6. Gentamycin (Sigma-Aldrich, UK).
7. 5-Fluorocytosine (Sigma-Aldrich, UK).
8. Blood agar base (Sigma-Aldrich, UK).
9. Agar (Sigma-Aldrich, UK).
10. Tryptone (Sigma-Aldrich, UK).
11. Sterile defibrinated rabbit blood.
13. Rubber caps from sodium heparin vacutainer tubes (#368480, Scientific Laboratory Supplies, UK).
14. Parafilm (VWR, UK).
15. Sterile broad forceps (Scientific Laboratory Supplies, UK).
16. Sterile Watchmakers’ forceps (Scientific Laboratory Supplies, UK).
17. Perspex dissection screen.
18. Sterile microscope slides (VWR, UK).
19. Sterile 13 mm microscope cover glasses (VWR, UK).
20. Sterile broad microspatula (Scientific Laboratory Supplies, UK).
21. Sterile 1.5 ml graduated microcentrifuge tubes (Anachem, UK).
22. Sterile 1 ml plastic Pasteur pipettes (Scientific Laboratory Supplies, UK).
23. 28 °C humidified incubator.
24. Inverted microscope.

2.2 Biological Cloning of T. cruzi

2.2.1 Micromanipulation

1. Blood agar base (Sigma-Aldrich, UK).
2. Agar (Sigma-Aldrich, UK).
3. Tryptone (Sigma-Aldrich, UK).
4. Sodium chloride, NaCl (Sigma-Aldrich, UK).
5. Sterile defibrinated rabbit blood.
6. Gentamycin (Sigma-Aldrich, UK).
7. 5-Fluorocytosine (Sigma-Aldrich, UK).
8. Mercuric chloride, HgCl₂ (Sigma-Aldrich, UK).
9. Hydrochloric acid sp.gr.1.18, HCl (VWR, UK).
10. Absolute ethanol (analytical reagent grade).
11. Sterile microcapillary tubes (Sigma-Aldrich, UK).
12. Bunsen burner (Scientific Laboratory Supplies, UK).
14. Sterile 13 mm microscope cover glasses (VWR, UK).
15. Sterile 7 ml Bijou tubes (Sterilin, UK).
17. Sterile Watchmakers’ forceps (Scientific Laboratory Supplies, UK).
18. 28 °C humidified incubator.
19. Inverted microscope.
2.2.2 Plating on Solid Medium

1. RPMI-1640 liquid medium (Sigma-Aldrich, UK #R0883).
2. Tryptone (Sigma-Aldrich, UK).
3. HEPES sodium salt (Sigma-Aldrich, UK).
4. Hemin (Sigma-Aldrich, UK).
5. Sodium hydroxide, NaOH (Sigma-Aldrich, UK).
6. Heat-inactivated fetal calf serum (Sigma-Aldrich, UK).
7. Sodium glutamate (Sigma-Aldrich, UK).
8. Sodium pyruvate (Sigma-Aldrich, UK).
9. Streptomycin (Sigma-Aldrich, UK).
10. Penicillin (Sigma-Aldrich, UK).
11. Blood agar base (Sigma-Aldrich, UK).
12. Agar (Sigma-Aldrich, UK).
13. Sterile defibrinated rabbit blood.
15. Low melting point (LMP) agarose (Sigma-Aldrich, UK).
16. Sodium chloride, NaCl (Sigma-Aldrich, UK).
17. Gentamycin (Sigma-Aldrich, UK).
18. Parafilm (VWR, UK).
19. Sterile 90 mm petri dishes (Sterilin, UK).
20. Sterile 200 μl pipette tips (Star Laboratories, UK).
21. Sterile 48-well cell culture plates (Becton Dickinson, USA).
22. 28 °C humidified incubator.
23. Inverted microscope.

2.2.3 Limiting Dilution

1. RPMI-1640 liquid medium (Sigma-Aldrich, UK #R0883).
2. Tryptone (Sigma-Aldrich, UK).
3. HEPES sodium salt (Sigma-Aldrich, UK).
4. Hemin (Sigma-Aldrich, UK).
5. Sodium hydroxide, NaOH (Sigma-Aldrich, UK).
6. Heat-inactivated fetal calf serum (Sigma-Aldrich, UK).
7. Sodium glutamate (Sigma-Aldrich, UK).
8. Sodium pyruvate (Sigma-Aldrich, UK).
9. Streptomycin (Sigma-Aldrich, UK).
10. Penicillin (Sigma-Aldrich, UK).
12. Sterile 96-microwell culture plates (Nunc, UK).
13. 28 °C humidified incubator.
2.3 Preparation of Parasite Genomic DNA

1. Gentra Puregene tissue kit (Qiagen, UK).
2. High Pure PCR template preparation kit (Roche, UK).
3. DNAzol® solution (Life Technologies, UK).
5. Microcentrifuge.
7. Water bath.
8. Phosphate-buffered saline (PBS) (Sigma-Aldrich, UK).
9. Absolute isopropanol (analytical reagent grade).
10. Absolute ethanol (analytical reagent grade).
11. Sodium hydroxide, NaOH (Sigma-Aldrich, UK).
12. Sterile 15 ml centrifuge tubes (Greiner Bio-One, UK).
13. Sterile 1.5 ml graduated microcentrifuge tubes (Anachem, UK).

2.4 PCR-RFLP Amplification

1. Oligonucleotides to amplify the D7 divergent domain of the 24Sα rRNA gene (LSU rDNA), heat shock protein 60 (HSP60), and glucose-6-phosphate isomerase (GPI) (see Table 1).
2. T. cruzi genomic DNA.
3. 10× NH₄ buffer (Bioline, UK).
4. 50 mM MgCl₂ solution (Bioline, UK).
5. Deoxynucleotide solution mix (10 mM stock of each dNTP) (New England Biolabs, UK).
6. BIOTAQ™ DNA polymerase (Bioline, UK).

Table 1
PCR-RFLP gene fragments and primer details

<table>
<thead>
<tr>
<th>PCR-RFLP target</th>
<th>Primer name</th>
<th>Primer Sequence (5′ → 3′)</th>
</tr>
</thead>
<tbody>
<tr>
<td>LSU rDNA&lt;sup&gt;a&lt;/sup&gt;</td>
<td>D71</td>
<td>AAGGTGCGTCGACAGTGTGG (20)</td>
</tr>
<tr>
<td>LSU rDNA&lt;sup&gt;a&lt;/sup&gt;</td>
<td>D72</td>
<td>TTTTCAGAATGGCCGAACAGT (21)</td>
</tr>
<tr>
<td>HSP60&lt;sup&gt;b&lt;/sup&gt;</td>
<td>HSP60_for</td>
<td>GTGGTATGGGTGACATGTAC (20)</td>
</tr>
<tr>
<td>HSP60&lt;sup&gt;b&lt;/sup&gt;</td>
<td>HSP60_rev</td>
<td>CGAGCAGCAGAGCGAAACAT (20)</td>
</tr>
<tr>
<td>GPI&lt;sup&gt;c&lt;/sup&gt;</td>
<td>GPI_for</td>
<td>GCCATGTGAAGCTTTGAGGCCTTTTTCAG (29)</td>
</tr>
<tr>
<td>GPI&lt;sup&gt;c&lt;/sup&gt;</td>
<td>GPI_rev</td>
<td>TGTAAGGGCCCAGTGAGAGCGTTCGTTGAATAGC (34)</td>
</tr>
</tbody>
</table>

<sup>a</sup> Primer sequences according to Brisse et al. [73]
<sup>b</sup> Primer sequences according to Strurm et al. [74]
<sup>c</sup> Primer sequences according to Gaunt et al. [75]
7. Sterile 0.2 ml 96-well PCR reaction plates and adhesive plate seals (Fisher Scientific, UK) or 0.2 ml PCR tube strips and caps (VWR, UK).
8. PCR machine.
10. Sterile 1.5 ml graduated microcentrifuge tubes (Anachem, UK).

2.5 Nuclear MLST PCR Amplification

1. Oligonucleotides to amplify 3-hydroxy-3-methylglutaryl-CoA reductase (HMCOAR), glucose-6-phosphate isomerase (GPI), mitochondrial peroxidase (TcMPX), and rho-like GTP binding protein (RHOI) (see Table 2).
2. *T. cruzi* genomic DNA.
3. 5× colorless GoTaq® reaction buffer (Promega, UK).
4. Deoxynucleotide solution mix (10 mM stock of each dNTP) (New England Biolabs, UK).
5. GoTaq® DNA polymerase (Promega, UK).
6. Sterile 0.2 ml 96-well PCR reaction plates and adhesive plate seals (Fisher Scientific, UK) or 0.2 ml PCR tube strips and caps (VWR, UK).
7. PCR machine.
8. Microcentrifuge.
9. Sterile 1.5 ml graduated microcentrifuge tubes (Anachem, UK).

2.6 Maxicircle MLST PCR Amplification

1. Oligonucleotides to amplify ten maxicircle gene fragments (see Table 3 and Fig. 2).
2. *T. cruzi* genomic DNA.
3. 10× NH₄ buffer (Bioline, UK).
4. 50 mM MgCl₂ solution (Bioline, UK).
5. Deoxynucleotide solution mix (10 mM stock of each dNTP) (New England Biolabs, UK).
6. BIOTAQ™ DNA polymerase (Bioline, UK).
7. Sterile 0.2 ml 96-well PCR reaction plates and adhesive plate seals (Fisher Scientific, UK) or 0.2 ml PCR tube strips and caps (VWR, UK).
8. PCR machine.
10. Sterile 1.5 ml graduated microcentrifuge tubes (Anachem, UK).

2.7 MLMT PCR Amplification

1. Oligonucleotides to amplify 28 microsatellite loci. Five fluorescent dyes with different emission spectra are used to label the forward primers: 6-FAM and TET (Proligo, Germany) and NED, PET, and VIC (Applied Biosystems, UK) (see Table 4 and Fig. 3).
<table>
<thead>
<tr>
<th>Gene fragment</th>
<th>Primer name</th>
<th>Primer sequence (5’ → 3’)</th>
<th>Annealing temperature (°C)</th>
<th>Amplicon size (bp)</th>
<th>Sequence start 5’</th>
<th>Sequence start 3’</th>
<th>Sequenced fragment (bp)</th>
</tr>
</thead>
<tbody>
<tr>
<td>HMCOAR</td>
<td>HMCOAR Fwd</td>
<td>AGGAGGCTTTTGTAGTCCACA (20)</td>
<td>55</td>
<td>564</td>
<td>TGAGTCCA</td>
<td>TCCAACA</td>
<td>554</td>
</tr>
<tr>
<td></td>
<td>HMCOAR Rvs</td>
<td>TCCAACAACACCAACCTCAA (20)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GPI</td>
<td>GPI Fwd</td>
<td>CGCCATGTTGTGAATATTGG (20)</td>
<td>55</td>
<td>424</td>
<td>TGAATATT</td>
<td>CAATGAGT</td>
<td>405</td>
</tr>
<tr>
<td></td>
<td>GPI Rvs</td>
<td>GGCGGACCACAATGAGTATC (20)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>TcMPX</td>
<td>TcMPX Fwd</td>
<td>ATGTTCGTCGTATGGCC (18)</td>
<td>55</td>
<td>678</td>
<td>TACATGGA</td>
<td>CGCACCGT</td>
<td>505</td>
</tr>
<tr>
<td></td>
<td>TcMPX Rvs</td>
<td>TGCGTTTTTCTCAAAAATATTC (21)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>RHO1</td>
<td>RHO1 Fwd</td>
<td>AGTTGCTGCTTCCCATCAAT (20)</td>
<td>55</td>
<td>463</td>
<td>CTTCACAT</td>
<td>TCTGCACA</td>
<td>455</td>
</tr>
<tr>
<td></td>
<td>RHO1 Rvs</td>
<td>CTGCACAGTGATGCCTGCT (20)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
### Table 3
Maxicircle MLST gene fragments and primer details

<table>
<thead>
<tr>
<th>Gene fragment</th>
<th>Genome position</th>
<th>Primer name</th>
<th>Primer sequence (5’ → 3’)</th>
<th>Annealing temperature (°C)</th>
<th>Amplicon Size (bp)</th>
<th>Sequence Start 5’</th>
<th>Sequence Start 3’</th>
<th>Sequenced Fragment (bp)</th>
</tr>
</thead>
<tbody>
<tr>
<td>12S rRNA</td>
<td>639-901</td>
<td>12S Fwd</td>
<td>GTTTATTAAATGCCTTTGTCAAGAA (26)</td>
<td>50</td>
<td>299</td>
<td>GTCTAAGA</td>
<td>TACGTATT</td>
<td>263</td>
</tr>
<tr>
<td></td>
<td></td>
<td>12S Rvs</td>
<td>GCCCCAATACACGACAAA (19)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>9S rRNA</td>
<td>1077-1309</td>
<td>9S Fwd</td>
<td>TGCAATTGCCTTAGGTTGTA (21)</td>
<td>50</td>
<td>302</td>
<td>TAAAATCG</td>
<td>TATTATTA</td>
<td>233</td>
</tr>
<tr>
<td></td>
<td></td>
<td>9S Rvs</td>
<td>TCCACACCAATAGCAGCTG (22)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CYT b</td>
<td>4126-4733</td>
<td>Sp18 Fwd</td>
<td>GACAGGATTGAGAAAGCGAGAGAG (23)</td>
<td>50</td>
<td>717</td>
<td>TTTGTYYT</td>
<td>TAATAYCA</td>
<td>608</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Sp18 Rvs</td>
<td>CAAACCTATCACAAAAAGCATCTG (24)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Murfla</td>
<td>6011-6393</td>
<td>Murfla Fwd</td>
<td>AAGGCRATGGRATAGWRCCTATAC (25)</td>
<td>50</td>
<td>482</td>
<td>ACTAAGYA</td>
<td>ACTTTYTA</td>
<td>383</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Murfla Rvs</td>
<td>TGGAAACATATATACAGATGRGGGA (26)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Murflb</td>
<td>6528-6900</td>
<td>Murflb Fwd</td>
<td>ACMCCATCATCTTCCTCR (18)</td>
<td>50</td>
<td>423</td>
<td>CAAAATTT</td>
<td>GGATTTAT</td>
<td>373</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Murflb Rvs</td>
<td>CTTTGATYTATTGTGATTAACRKT (25)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ND1</td>
<td>7643-8011</td>
<td>ND1 Fwd</td>
<td>GCACCTTCTGAAATAATCGAAA (23)</td>
<td>50</td>
<td>400</td>
<td>TCGAAAAA</td>
<td>TTGTTAGC</td>
<td>369</td>
</tr>
<tr>
<td></td>
<td></td>
<td>ND1 Rvs</td>
<td>TAAACITCATCGGATGGTCC (25)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>COII</td>
<td>8194-8610</td>
<td>COII Fwd</td>
<td>GTTATTATCTTCTGTTGTTGTGTTGTTG (27)</td>
<td>50</td>
<td>560</td>
<td>CTTCTCTAC</td>
<td>ACCTRCCY</td>
<td>417</td>
</tr>
<tr>
<td></td>
<td></td>
<td>COII Rvs</td>
<td>AACAATTTCGATAATCCATGT (22)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ND4</td>
<td>12153-12392</td>
<td>ND4 Fwd</td>
<td>TTTTGGAAAGTGCTATTTTCCCA (23)</td>
<td>50</td>
<td>302</td>
<td>AATTTTAA</td>
<td>CGGYRTC</td>
<td>240</td>
</tr>
<tr>
<td></td>
<td></td>
<td>ND4 Rvs</td>
<td>CTTCAACATGCATTCTCCGGT (21)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ND5a</td>
<td>13829-14250</td>
<td>ND5a Fwd</td>
<td>TATGRTACYTTTTTCATGTCRG (24)</td>
<td>50</td>
<td>503</td>
<td>GTACATAY</td>
<td>TYTGYTA</td>
<td>422</td>
</tr>
<tr>
<td></td>
<td></td>
<td>ND5a Rvs</td>
<td>GTCCCTCCATYGCATTYCGG (19)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ND5b</td>
<td>14274-14640</td>
<td>ND5b Fwd</td>
<td>ARAGTACACAGTGTGGGTRFCAYA (24)</td>
<td>50</td>
<td>444</td>
<td>TGATTRCC</td>
<td>GYARACCA</td>
<td>367</td>
</tr>
<tr>
<td></td>
<td></td>
<td>ND5b Rvs</td>
<td>CTTGGYARATACACCGACAAA (21)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*Genome position according to the TcI Sylvio X10/1 reference maxicircle genome [76]

*Amplicon size according to TcI Sylvio X10/1. Indels in other strains may cause size variation

*Sequence length according to TcI Sylvio X10/1. Indels in other strains may cause length variation
2. 0.5× TE buffer (10 mM Tris–HCl and 1 mM EDTA (pH 8)) (both Sigma-Aldrich, UK).

3. *T. cruzi* genomic DNA.

4. 10× ThermoPol reaction buffer (New England Biolabs, UK).

5. 50 mM MgCl₂ solution (Bioline, UK).

6. Deoxynucleotide solution mix (10 mM stock of each dNTP) (New England Biolabs, UK).


---

**Fig. 2** Maxicircle primer positions in 96-well plate
<table>
<thead>
<tr>
<th>Chromosome&lt;sup&gt;a&lt;/sup&gt;</th>
<th>Primer code</th>
<th>Repeat type</th>
<th>Forward/reverse primer (5′–3′)</th>
</tr>
</thead>
<tbody>
<tr>
<td>6</td>
<td>6529(CA)&lt;sup&gt;a&lt;/sup&gt;</td>
<td>(CA)&lt;sub&gt;n&lt;/sub&gt;</td>
<td>TGTGAAATGATTTGACCCCGA AGAGTCGCCCGCAAAAGTAT</td>
</tr>
<tr>
<td>6</td>
<td>6529(TA)&lt;sup&gt;b&lt;/sup&gt;</td>
<td>(TA)&lt;sub&gt;n&lt;/sub&gt;</td>
<td>TGAAGGAGATTCTCTGCGGT CTCTCATCTTTTTGTGTGTCGG</td>
</tr>
<tr>
<td>6</td>
<td>mclf10</td>
<td>(CA)&lt;sub&gt;n&lt;/sub&gt;(A(CA)&lt;sub&gt;n&lt;/sub&gt;)</td>
<td>GCGTAGCGATTCATTTCC ATCCGCTACACTACCAC</td>
</tr>
<tr>
<td>10</td>
<td>6855(TA)(GA)</td>
<td>(TA)&lt;sub&gt;n&lt;/sub&gt;(GA)&lt;sub&gt;n&lt;/sub&gt;</td>
<td>TGTGATCAACCGCGCATAAAAT TTCCATTGCCTGTTTTTAA</td>
</tr>
<tr>
<td>15</td>
<td>11863(CA)</td>
<td>(CA)&lt;sub&gt;n&lt;/sub&gt;</td>
<td>AGTTGACATTCCCACAGCAAG CCCTGATGCGCAGACTTCTT</td>
</tr>
<tr>
<td>19</td>
<td>TcUn3</td>
<td>Unknown</td>
<td>CTAAAGAGATACAAAGAGGGAAGG CTGTTATTTCAATAACAGGGG</td>
</tr>
<tr>
<td>19</td>
<td>10101(TA)</td>
<td>(TA)&lt;sub&gt;n&lt;/sub&gt;</td>
<td>AACCCGCGCAAGATACATTAG TCCATTTGCGACACACACA</td>
</tr>
<tr>
<td>24</td>
<td>8741(TA)</td>
<td>(TA)&lt;sub&gt;n&lt;/sub&gt;</td>
<td>TGTAAACGCTAGGTCTCAAACTCC TTGCATTGCGGAGCCAAAC</td>
</tr>
<tr>
<td>27</td>
<td>10101(TC)</td>
<td>(TC)&lt;sub&gt;n&lt;/sub&gt;</td>
<td>CGTACGACGCGACACAAAC ACAAGTGGTGGGACGACAAAG</td>
</tr>
<tr>
<td>27</td>
<td>10101(CA)&lt;sup&gt;c&lt;/sup&gt;</td>
<td>(CA)&lt;sub&gt;n&lt;/sub&gt;</td>
<td>GTGTCGTTGCTGCCCAAACTC AACTTGCCAAATGTTAGGG</td>
</tr>
<tr>
<td>27</td>
<td>10101(CA)&lt;sup&gt;a&lt;/sup&gt;</td>
<td>(CA)&lt;sub&gt;n&lt;/sub&gt;</td>
<td>GTCGCCATCATGTACAAACG CTGTGCGGAATGTCGTCAA</td>
</tr>
<tr>
<td>34</td>
<td>6559(TC)</td>
<td>(TC)&lt;sub&gt;n&lt;/sub&gt;</td>
<td>CGCTCCTAAGGACACCTTAC ATATGGACGCGTGGAGGTGC</td>
</tr>
<tr>
<td>37</td>
<td>10187(TTA)</td>
<td>(TTA)&lt;sub&gt;n&lt;/sub&gt;</td>
<td>GAGAGAGATTCCGAAAATCATTAGC CATGTCCTTCTCCTCGTAAA</td>
</tr>
<tr>
<td>37</td>
<td>10187(CA)(TA)</td>
<td>(CA)&lt;sub&gt;n&lt;/sub&gt;(TA)&lt;sub&gt;n&lt;/sub&gt;</td>
<td>CATGTCAATAGTGGCCACAC GCACATGTGGTTGGTGGAA</td>
</tr>
<tr>
<td>37</td>
<td>10187(TA)</td>
<td>(TA)&lt;sub&gt;n&lt;/sub&gt;</td>
<td>AGAAAAAGTTTACAAACGACGC CGATGGAGACGAGAAAACAA</td>
</tr>
<tr>
<td>37</td>
<td>10187(GA)</td>
<td>(GA)&lt;sub&gt;n&lt;/sub&gt;</td>
<td>GTCACACACTACGCATGACA ACTGCACATACCCCCTTTG</td>
</tr>
</tbody>
</table>

(continued)
8. Sterile 0.2 ml 96-well PCR reaction plates (Fisher Scientific, UK) and adhesive plate seals or 0.2 ml PCR tube strips and caps (VWR, UK).
9. PCR machine.
10. Microcentrifuge.
11. Sterile 1.5 ml graduated microcentrifuge tubes (Anachem, UK).

### 2.8 Agarose Gel Electrophoresis

1. Molecular grade agarose (Bioline, UK).
2. NuSieve™ GTG™ agarose (Lonza, UK).
3. 1× TAE buffer (40 mM Tris–HCl, 20 mM acetic acid, and 1 mM EDTA (pH 8)) (Sigma-Aldrich, UK).
4. 10 mg/ml ethidium bromide (Sigma-Aldrich, UK) (see Note 3).
5. Molecular weight ladders: Hyperladder™ IV and V (Bioline, UK).
6. 5× DNA loading buffer blue (Bioline, UK).

<table>
<thead>
<tr>
<th>Chromosome</th>
<th>Primer code</th>
<th>Repeat type</th>
<th>Forward/reverse primer (5′–3′)</th>
</tr>
</thead>
<tbody>
<tr>
<td>37</td>
<td>TcUn2</td>
<td>Unknown</td>
<td>AACAAAATCTAGCGTCTACCATCC</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>GGTGTTGGCGTGTATGATTTG</td>
</tr>
<tr>
<td>39</td>
<td>6925(TG)b</td>
<td>(TG)ₙ</td>
<td>GAAACGCACCTACCCACAC</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>GGTAGCAAACGCCAAACTTTC</td>
</tr>
<tr>
<td>39</td>
<td>7093(TC)</td>
<td>(TC)ₙ</td>
<td>CCAACATTCACAAAGGGAAAA</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>GATGAATAGTGCCGGATCT</td>
</tr>
<tr>
<td>39</td>
<td>6925(CT)</td>
<td>(CT)ₙ</td>
<td>CATCAAGGAAAAACCGAGGA</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>CGGTACCACCTCAAGGAAGA</td>
</tr>
<tr>
<td>39</td>
<td>7093(TA)c</td>
<td>(TA)ₙ</td>
<td>CGTGTGCACAGGAGAGAAAA</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>CGTTTGGAGGAGGATTGAGA</td>
</tr>
<tr>
<td>39</td>
<td>6925(TG)a</td>
<td>(TG)ₙ</td>
<td>TCGTTCTCTTTACGCTTGCA</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>TAGCAGCACCACAAACAAAAAG</td>
</tr>
<tr>
<td>39</td>
<td>7093(TCC)</td>
<td>(TCC)ₙ</td>
<td>AGACGTTTATATTGCAGCC</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>AGCCACATCCACATTTCC</td>
</tr>
<tr>
<td>40</td>
<td>11283(TCG)</td>
<td>(TCG)ₙ</td>
<td>ACCACAGGAGGACATGAAG</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>TGTCAGGAAACAGCGAAG</td>
</tr>
<tr>
<td>40</td>
<td>11283(TA)b</td>
<td>(TA)ₙ</td>
<td>AACACTCCTCCACCTCACAGG</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>TTTGAATGCGAGGTGTACA</td>
</tr>
</tbody>
</table>

*Chromosomal assignment based on Weatherly et al. [79]*
7. Gel electrophoresis equipment (e.g., Jencons midi-horizontal gel electrophoresis system with 16-well combs and 13 × 15 cm casting trays) and power pack.

8. Microwave.

9. UV transilluminator.

2.9 PCR Purification

1. QIAquick PCR purification kit (Qiagen, UK).

2. Absolute ethanol (analytical reagent grade).

3. Absolute isopropanol (analytical reagent grade).

4. 0.5× TE buffer (10 mM Tris–HCl and 1 mM EDTA (pH 8)) (both Sigma-Aldrich, UK).

5. Microcentrifuge.
6. Sterile 1.5 ml graduated microcentrifuge tubes (Anachem, UK).
7. Sterile 0.5 ml graduated microcentrifuge tubes (Anachem, UK).

2.10 Restriction Enzyme Digestions

1. EcoRV restriction endonuclease, corresponding 10× buffer and 100× bovine serum albumin (BSA) (New England Biolabs, UK).
2. HhaI restriction endonuclease, corresponding 10× buffer and 100× BSA (New England Biolabs, UK).
3. Microcentrifuge.
4. Sterile 1.5 ml graduated microcentrifuge tubes (Anachem, UK).
5. 37 °C incubator.

2.11 Dye Terminator DNA Sequencing

2. PCR machine.
3. Absolute ethanol (analytical reagent grade).
4. 96-well optical reaction plates with barcodes (Applied Biosystems, UK).
5. Hi-Di™ deionized formamide (Applied Biosystems, UK).
6. Refrigerated centrifuge.
7. Vortex.
8. 16-Capillary 3730 DNA Analyzer (Applied Biosystems, UK) (see Note 4).

2.12 MLMT PCR Product Multiplexing and Allele Size Determination

1. 96-well optical reaction plates with barcodes (Applied Biosystems, UK).
2. GeneScan™-500 LIZ™ size standard (Applied Biosystems, UK).
3. Hi-Di™ deionized formamide (Applied Biosystems, UK).
4. 16-Capillary 3730 DNA Analyzer (Applied Biosystems, UK).

3 Methods

3.1 Isolation of T. cruzi

3.1.1 Direct Hemoculture from Patients/Mammals

1. Prepare biphasic 4 N (USMARU) culture medium by adding 4 % (w/v) blood agar base, 0.6 % (w/v) agar, 0.6 % (w/v) NaCl, and 0.5 % (w/v) tryptone (all Sigma-Aldrich, UK) to H₂O and dissolve by autoclaving (121 °C for 15 min). Cool the medium to 50 °C and aseptically add 10 % (v/v) sterile defibrinated rabbit blood, 150 μg/ml gentamycin, and 150 μg/ml 5-fluorocytosine (both Sigma-Aldrich, UK) (see Note 5).

2. Aliquot 2 ml of biphasic 4 N culture medium into the bottom of a sterile Nunclon™ Δ flat sided tube (Nunc, UK) and allow to set at an angle, forming a slope.
3. Once set, overlay each culture with 500 μl of 0.9 % sterile NaCl, containing 150 μg/ml gentamycin and 150 μg/ml 5-fluorocytosine.

4. Prepare liver infusion tryptose (LIT) medium by dissolving 25 g liver infusion broth (Difco™, Becton Dickinson, USA), 5 g tryptone, 4 g NaCl, 2 g glucose (Sigma-Aldrich, UK), 0.4 g KCl (Sigma-Aldrich, UK), and 3.15 g Na₂HPO₄ (Sigma-Aldrich, UK) in 900 ml H₂O and adjust the pH to 7.4. Autoclave (121 °C for 15 min) and cool the medium to 50 °C. Add 25 g hemin, dissolved in 1 ml 1 N NaOH and 100 ml heat-inactivated fetal calf serum (both Sigma-Aldrich, UK).

5. For adult human samples, extract 15 ml venous blood using a sterile 20 ml BD Plastipak™ syringe with needle (Becton Dickinson, USA) and transfer to a BD Vacutainer® plus plastic K₂ EDTA tube (Becton Dickinson, USA) to prevent coagulation.

6. If isolating from mammals, take 1–2 ml blood by cardiac puncture, using a sterile 5 ml BD Plastipak™ syringe with needle (or 1 ml/2 ml syringes for smaller animals), after anesthetising the mammal by intramuscular administration of ketamine hydrochloride (100 mg/kg body weight) (Sigma-Aldrich, UK) and sterilizing the thorax first with iodised 70 % (v/v) ethanol (5 g iodine/l) and then non-iodized 70 % (v/v) ethanol.

7. For patient samples, transfer blood into a sterile 15 ml centrifuge tube (Greiner Bio-One, UK) and centrifuge for 10 min at 1,200 × g and 4 °C.
   (a) Discard all but 0.5 ml plasma and packed red cells.
   (b) Add 8 ml LIT medium to packed red cells.
   (c) Centrifuge for 10 min at 1,200 × g and 4 °C.
   (d) Carefully discard the supernatant.
   (e) Resuspend in 6 ml LIT medium.
   (f) Aliquot 2 ml of packed red cells to three separate 4 N culture tubes.
   (g) Seal each tube with a rubber cap from a 10 ml vacutainer tube (Scientific Laboratory Supplies, UK) and secure with Parafilm (VWR, UK). Rubber caps must be autoclaved prior to use.
   (h) Incubate cultures at 28 °C for 3–6 months, depending on strain growth rate. Once logarithmic phase cells become microscopically visible, parasites can be seeded into supplemented RPMI-1640 axenic culture medium (as described in Subheading 3.2.2).
   (i) For long-term cryopreservation of parasites, supplement late logarithmic phase cultures with sterile 10 % glycerol (v/v) (VWR, UK) and prepare aliquots in sterile cryovials (Nunc, Denmark). Store cryovials at –70 °C for 24 hours, before transfer to liquid nitrogen.
8. For mammal samples, inoculate the blood directly into several 4 N culture tubes.

9. If biologically cloning directly from blood (as described in Subheading 3.2), leave the whole blood to settle for 1 h in the BD Vacutainer® EDTA tube or centrifuge at a low speed (40 × g for 5 min) and then incubate at 37 °C for 45 min to ensure motile trypomastigotes have dispersed throughout the plasma, prior to cloning.

10. If directly extracting parasite genomic DNA from blood (as described in Subheading 3.3.2), dilute packed red cells in guanidine-EDTA (6 M guanidine, 0.2 M EDTA) (Sigma-Aldrich, UK) at a 1:1 ratio and store at 4 °C.

Xenodiagnosis can be undertaken by feeding up to 10–20 uninfected colony-reared triatomine bugs (third or fourth nympha instars) on each suspected patient/mammal before isolating parasites after ~3 weeks as described below:

1. Prepare biphasic 4 N culture medium in sterile Nunclon™ flat sided tubes as described in Subheading 3.1.1.

2. Prepare White’s solution consisting of 0.025 g HgCl₂ (see Note 6), 0.65 g NaCl (both Sigma-Aldrich, UK), 0.125 ml conc. HCl (sp. gr. 1.18) (VWR, UK), 25 ml absolute ethanol, and 75 ml H₂O.

3. Immerse the bugs in White’s solution for 10 min, rinse in 0.9 % sterile NaCl containing 300 μg/ml gentamycin and 300 μg/ml 5-fluorocytosine and dry (all Sigma-Aldrich, UK).

4. Aseptically dissect the intestinal contents of each bug into sterile saline (containing 300 μg/ml gentamycin and 300 μg/ml 5-fluorocytosine) on a sterile microscope slide (VWR, UK), behind a protective screen in a Class II microbiological safety cabinet. Dissection can be performed by holding the bug upside down in a pair of broad forceps, then using a pair of watchmakers’ forceps (both Scientific Laboratory Supplies, UK) to pull the last abdominal segment away, extruding the gut onto a microscope slide.

5. Homogenize the intestinal contents using a sterile broad microspatula (Scientific Laboratory Supplies, UK) and discard the abdomen apex.

6. Remove the majority of intestinal homogenate from the dissection slide to a sterile 1.5 ml graduated microcentrifuge tube (Anachem, UK), using a sterile 1 ml plastic Pasteur pipette (Scientific Laboratory Supplies, UK) and place a sterile microscope cover glass over the remainder.

7. Examine slide microscopically and if parasites are observed, transfer 20 μl of inoculum to a 4 N culture tube.
3.2 Biological Cloning of T. cruzi

3.2.1 Biological Cloning of Parasites by Micromanipulation

Variations of this protocol, including different under- and over-lay media are published in full in [41]. We describe below a protocol which favors growth of *T. cruzi* strains from all DTUs:

1. Prepare biphasic 4 N culture medium as described in Subheading 3.1.1 but without gentamycin and 5-fluorocytosine.

2. Aliquot 2 ml of biphasic 4 N culture medium into the bottom of sterile 7 ml Bijou tubes (Sterilin, UK) and leave to set. Once set, overlay each culture with 750 μl of 0.9 % sterile NaCl, containing 100 μg/ml gentamycin (Sigma-Aldrich, UK) and 100 μg/ml 5-fluorocytosine (Sigma-Aldrich, UK).

3. Empirically prepare a dilute solution of logarithmic-phase *T. cruzi* epimastigotes (from axenic culture, patient blood or infected triatomine bug intestinal contents) such that microdrops delivered from microcapillaries contain a single parasite or no parasites (see Note 7).

4. Prepare fine microcapillaries by rotating a microcapillary tube (Sigma-Aldrich, UK) in a Bunsen flame, removing, and pulling apart the two ends to form a fine intervening microcapillary (each original microcapillary tube yields two microcapillaries).

5. On a microscope slide, place a sterile 13 mm microscope cover glass onto a small drop of sterile H₂O (for adhesion); dispense a microdrop of diluted culture onto the cover glass from a microcapillary tube and cover the drop with a second cover glass. Drops which occupy no more than one microscopic field at 400× magnification are ideal.

6. Microscopically examine the drop through multiple planes of vision, for the presence of parasites.

7. Transfer cover glass pairs with drops containing no organisms (control cultures) or a single parasite to 4 N cultures using sterile watchmakers’ forceps. Discard all microdrops which contain more than one parasite.

8. Incubate all cultures at 28 °C for 3–6 months, depending on strain growth rate. Discard the entire series if any of the control cultures become positive. Once logarithmic phase cells become microscopically visible, parasites can be seeded into supplemented RPMI-1640 axenic culture medium (as described in Subheading 3.2.2).
1. Prepare sterile stock solutions (100×) of tryptone (0.175 g/ml, autoclaved), HEPES (1 M, pH 7.2, filter-sterilized), and hemin (2.5 mg/ml in 0.01 M NaOH, autoclaved) (all Sigma-Aldrich, UK).

2. Supplement RPMI-1640 medium (Sigma-Aldrich, UK #R0883) with 0.5 % (w/v) tryptone, 20 mM HEPES buffer (pH 7.2), 30 mM hemin, 10 % (v/v) heat-inactivated fetal calf serum, 2 mM sodium glutamate, 2 mM sodium pyruvate, 250 μg/ml streptomycin, and 250 U/ml penicillin (all Sigma-Aldrich, UK). Filter-sterilize the glutamine/pyruvate/penicillin solution before use.

3. Prepare blood agar plates by adding 10.8 ml biphasic 4 N culture medium (with 100 μg/ml gentamycin and 100 μg/ml 5-fluorocytosine) as described in Subheading 3.1.1 to sterile 90 mm petri dishes (Sterilin, UK).


5. Mix 10^2–10^3 logarithmic phase cells with 2.4 ml (w/v) supplemented RPMI-1640 medium and 0.6 ml molten 3 % (w/v) LMP agarose containing 0.9 % NaCl (w/v) (all Sigma-Aldrich, UK).

6. Pour this overlay onto a blood agar plate and allow to set.

7. Seal plates with Parafilm (VWR, UK) to minimize evaporation and incubate at 28 °C in a humidified atmosphere of 5 % CO₂.

8. Once colonies become visible (after 3–6 months, depending on strain growth rate), examine microscopically and remove clones using sterile 200 μl pipette tips. Inoculate each colony into 1 ml supplemented RPMI-1640 medium in a 48-well cell culture plate (Becton Dickinson, USA).

3.2.3 Biological Cloning of Parasites by Limiting Dilution

1. Serially dilute logarithmic phase cells to achieve a final concentration of 0.5 parasites/ml in a total volume of 20 ml supplemented RPMI-1640 medium (as described in Subheading 3.2.2).

2. Aliquot 200 μl of dilute culture into each well of a sterile 96-microwell culture plate (Nunc, UK).

3. Examine each well microscopically and mark those containing single organisms.

4. Seal each plate with Parafilm (VWR, UK) and incubate at 28 °C in an atmosphere of 5 % CO₂.

5. After 4–8 weeks, expand marked wells with sufficient numbers (∼10^6/ml) of dividing cells into larger axenic culture volumes.

3.3 Preparation of Parasite Genomic DNA

3.3.1 Parasite Genomic DNA Extraction from Epimastigote Culture

Extraction of genomic DNA from 10 ml epimastigote cultures can be achieved using a Gentra Puregene tissue kit (Qiagen, UK), according to a modified version of the manufacturer’s protocol (see Note 8). Additional necessary reagents are PBS (Sigma-Aldrich, UK), absolute isopropanol, and absolute ethanol. Cell
lysis buffer, protein precipitation solution, and DNA hydration solution are all stored at room temperature. Proteinase K and RNase A are both stored at 4 °C. The modified manufacturer's protocol is as follows:

1. Centrifuge 10 ml of late log phase culture (~10^7–10^8 trypanosomes) in a sterile 15 ml centrifuge tube (Greiner Bio-One, UK) at 800 × g for 10 min.
2. Discard the supernatant by inverting tubes onto absorbent paper and resuspend fully in PBS, then centrifuge again as previously.
3. Resuspend in 3 ml cell lysis buffer (incubate at 37 °C and/or vortex to remove clumps, if necessary).
4. Cell suspensions are now stable and can be stored at −20 °C for 1–2 weeks.
5. Add 15 μl proteinase K solution (100 μg/ml) and incubate at 55 °C for 1 h, inverting periodically.
6. Leave to cool to room temperature.
7. Add 15 μl RNase A solution (20 μg/ml), invert 25 times and incubate at 37 °C for 15–60 min.
8. Cool on ice for 3 min and then add 1 ml protein precipitation solution (room temperature).
9. Vortex tubes vigorously for 20 s and then centrifuge at 2,000 × g for 10 min (ensure a tight pellet forms).
10. Remove the supernatant and transfer to a new sterile 15 ml centrifuge tube.
11. Precipitate DNA by the addition of 3 ml absolute isopropanol (room temperature) and invert 50 times.
12. Centrifuge at 2,000 × g for 3 min and discard the supernatant by inverting tubes onto absorbent paper.
13. Wash the DNA pellet in 3 ml 70 % (v/v) ethanol (room temperature), invert 10 times and centrifuge at 2,000 × g for 1 min.
14. Carefully remove the supernatant by inverting tubes and draining onto absorbent paper.
15. Air-dry the DNA pellet with tubes inverted at an angle for a maximum of 15 min.
16. Resuspend the DNA pellet in 250 μl DNA hydration solution, incubate at 65 °C for 1–2 h and then at room temperature overnight.
17. Estimate the DNA yield by spectrophotometry. Successful DNA extractions will yield 100 ng/μl or more and an A260/280 of 1.8–2.0.
18. Store extracted genomic DNA at −20 °C.
Extraction of genomic DNA from clinical hemocultures can be achieved using a High Pure PCR template preparation kit (Roche, UK), according to the manufacturer’s protocol. Additional necessary reagents are PBS (Sigma-Aldrich, UK), absolute isopropanol, and absolute ethanol. Tissue lysis buffer, binding buffer, inhibitor removal buffer, wash buffer, and elution buffer are all stored at room temperature. Add absolute ethanol to the inhibitor removal buffer and the wash buffer, as instructed. Proteinase K is stored at 4 °C. Before beginning the DNA extraction, preheat the elution buffer to 70 °C. The manufacturer’s protocol is as follows:

1. To a sterile 1.5 ml graduated microcentrifuge tube (Anachem, UK) mix 200 µl sample material (1:1 blood/guanidine-EDTA) with 600 µl binding buffer and 100 µl Proteinase K and incubate at 70 °C for 10 min.
2. Add 200 µl absolute isopropanol and mix well by vortexing.
3. Apply 550 µl to a High Pure filter tube and centrifuge at 8,000 × g for 1 min.
4. Discard the flow-through.
5. Repeat steps 3 and 4 using the same High Pure filter tube.
6. Add 500 µl inhibitor removal buffer and centrifuge at 8,000 × g for 1 min.
7. Discard the flow-through.
8. Add 500 µl wash buffer and centrifuge at 8,000 × g for 1 min.
9. Discard the flow-through.
10. Repeat steps 7 and 8.
11. Centrifuge at 13,000 × g for 10 s.
12. Place the High Pure filter tube in a clean 1.5 ml graduated microcentrifuge tube.
13. Add 200 µl pre-warmed elution buffer and centrifuge at 8,000 × g for 1 min.
14. Estimate the DNA yield by spectrophotometry. Successful DNA extractions will yield 3 ng/µl or more and an A260/280 of 1.8–2.0.
15. Store extracted genomic DNA at −20 °C.

DNAzol® solution (Life Technologies, UK) can be used to extract *T. cruzi* genomic DNA from triatomine bug feces, following hind-gut dissection. Store DNAzol® solution, absolute ethanol, and NaOH at room temperature.

1. Lyse 50–100 µl of triatomine bug intestinal homogenate in a sterile 1.5 ml graduated microcentrifuge tube (Anachem, UK) by the addition of 1 ml DNAzol® solution.
2. Invert twice and incubate at room temperature for 3 min.
3. Precipitate DNA by the addition of 0.5 ml absolute ethanol (room temperature).
4. Pellet DNA by centrifuging at $13,000 \times g$ for 4 min.
5. Discard the supernatant and wash twice with 1 ml 70 % (v/v) ethanol ensuring not to disturb the pellet.
6. Resuspend the DNA pellet in 50 μl 8 mM NaOH (Sigma-Aldrich, UK).
7. Estimate the DNA yield by spectrophotometry. Successful DNA extractions will yield 100 ng/μl or more and an A260/280 of 1.8–2.0.
8. Store extracted genomic DNA at $-20 \, ^\circ\text{C}$.

### 3.4 PCR-RFLP

#### 3.4.1 PCR Amplification

1. Amplify the 24Sα rRNA (LSU rDNA) in a standard reaction containing: 1x NH$_4$ reaction buffer, 1.5 mM MgCl$_2$ (Bioline, UK), 0.2 mM dNTPs (New England Biolabs, UK), 1 pmol/μl of D71 and D72 primers (see Table 1), 1 U BIOTAQ™ DNA polymerase (Bioline, UK), and 10–100 ng of T. cruzi genomic DNA, made up to a total volume of 25 μl.
2. Reaction conditions for the 24Sα rRNA (LSU rDNA) are an initial denaturation step of 94 °C for 3 min and then 27 amplification cycles (94 °C for 1 min, 60 °C for 1 min, 72 °C for 1 min), followed by a final elongation step at 72 °C for 5 min.

3. Amplify both HSP60 and GPI in a standard reaction containing: 1x NH$_4$ reaction buffer, 2 mM MgCl$_2$ (Bioline, UK), 0.2 mM dNTPs (New England Biolabs, UK), 1 pmol/μl of HSP60_for and HSP60_rev primers (for HSP60) or GPI_for and GPI_rev (for GPI) (see Table 1), 1 U BIOTAQ™ DNA polymerase (Bioline, UK), and 10–100 ng of T. cruzi genomic DNA, made up to a total volume of 25 μl.

4. Reaction conditions for both HSP60 and GPI use a touch-down PCR strategy comprising an initial denaturation step of 3 min at 94 °C, followed by four cycles (94 °C for 30 s, 64 °C for 30 s, 72 °C for 1 min), followed by 28 cycles (94 °C for 30 s, 60 °C for 30 s, 72 °C for 1 min), and then a final elongation step at 72 °C for 10 min.

#### 3.4.2 Agarose Gel Electrophoresis

1. Visualize 10 μl of each 24Sα rRNA PCR product by gel electrophoresis using 3.5 % NuSieve™ GTG™ agarose gels (Lonza, UK) containing 0.5 μg/ml ethidium bromide (Sigma-Aldrich, UK) (see Note 9).

2. Visualize 5 μl of each HSP60 and GPI PCR product by gel electrophoresis using 1.5 % agarose gels (Bioline, UK) containing 0.5 μg/ml ethidium bromide.

3. Load samples into gel wells with 1 μl of 5x DNA loading buffer (Bioline, UK) and run 5 μl of Hyperladder™ V (for 24Sα
rRNA) or IV (for HSP60 and GPI) (Bioline, UK) as a molecular weight marker.

4. Run all gels at 100 V for 1–2 h in 1× TAE buffer and visualize under UV illumination, ensuring that the user is protected from the light source behind a UV shield.

5. If necessary, prior to restriction digestion, purify HSP60 and GPI PCR products using a QIAquick PCR purification kit (Qiagen, UK) to remove nonspecific products, as described in Subheading 3.4.3.

3.4.3 PCR Purification

Purification of all PCR products can be achieved using a QIAquick PCR purification kit (Qiagen, UK) with spin columns to remove contaminating primers, nucleotides, DNA polymerases etc. (see Note 10) All of the necessary reagents are included within the kit (add ethanol to buffer PE as instructed) and are stored at room temperature. The manufacturer’s protocol is as follows:

1. Add 5 volumes of buffer PB to 1 volume of the PCR reaction and mix.
2. Apply the sample to a QIAquick spin column placed in a 2 ml collection tube and centrifuge at >13,000×g for 30–60 s.
3. Discard the flow-through.
4. Add 0.75 ml of buffer PE (with ethanol added) to the QIAquick column and centrifuge at >13,000×g for 30–60 s.
5. Discard the flow-through and recentrifuge for 1 min at maximum speed.
6. Place the QIAquick column in a clean 1.5 ml graduated microcentrifuge tube (Anachem, UK).
7. To elute the DNA, add between 30 and 50 μl of buffer EB (10 mM Tris-Cl, 1 mM EDTA (pH 8)) to the center of the QIAquick membrane, incubate for 1–5 min and then centrifuge at >13,000×g for 1 min (see Note 11).
8. Purified PCR products can be stored at −20 °C until required.

3.4.4 Restriction Enzyme Digestion

1. Digest 10 μl of HSP60 or GPI PCR products (typically ~1 μg) in a reaction containing 0.25 U/μl of EcoRV or HhaI restriction endonucleases (New England Biolabs, UK), 100 ng/μl BSA and 1× quantity of the manufacturer’s recommended reaction buffer in a total volume of 20 μl.
2. Incubate reactions at 37 °C for 4 h.

3.4.5 Restriction Fragment Length Polymorphism Analysis

1. Visualize 10 μl of each reaction using either 1.5 % (GPI/HhaI) or 3 % agarose gels (HSP60/EcoRV) (Bioline, UK) containing 0.5 μg/ml ethidium bromide (Sigma-Aldrich, UK) (see Note 12).
2. Load samples into agarose wells with 1 μl of 5× DNA loading buffer (Bioline, UK) and run 5 μl of Hyperladder™ V (for
24Sα rRNA), IV (HSP60), or I (GPI) (Bioline, UK) as a molecular weight ladder.

3. Run all gels at 100 V for 1–2 h in 1× TAE buffer and visualize under UV illumination, ensuring that the user is shielded from the light source (see Note 13).

4. The genotype assignment system based on the number and size of the restriction fragment bands is shown in Table 5, Figs. 4 and 5. For additional details please refer to Lewis et al. [47].

### 3.5 Nuclear MLST

#### 3.5.1 PCR Amplification

1. Amplify each MLST target (HMCOAR, GPI, TcMPX and RHO1) in a standard reaction containing: 10 µl 5× colorless GoTaq® reaction buffer (Promega, UK), 0.2 mM dNTPs (New England Biolabs, UK), 0.2 µM of respective forward and reverse primers (see Table 2), 1 U GoTaq® DNA polymerase (Promega, UK), and 10–100 ng of T. cruzi genomic DNA, made up to a total volume of 50 µl.

2. Reaction conditions for all targets are an initial denaturation step of 94 °C for 5 min and then 35 amplification cycles (94 °C for 1 min, 55 °C for 1 min, 72 °C for 1 min), followed by a final elongation step at 72 °C for 5 min.

#### 3.5.2 Agarose Gel Electrophoresis

1. Visualize 5 µl of each PCR product by gel electrophoresis using 1.5 % agarose gels (Bioline, UK), as described in Subheading 3.4.2.

#### 3.5.3 PCR Purification

1. Purify all PCR products, as described in Subheading 3.4.3.

#### 3.5.4 Dye Terminator DNA Sequencing

Bidirectional sequencing can be performed using a BigDye™ Terminator v3.1 Cycle Sequencing kit (Applied Biosystems, UK). All of the necessary reagents are included within the kit. Big Dye Sequencing RR-100 is stored at −20 °C and sequencing buffer is stored at 4 °C (see Note 14). The modified version of the manufacturer’s protocol is as follows:

1. Use 0.5–2 µl of PCR reaction template (~5–20 ng) in a standard reaction containing 0.5 µl Big Dye sequencing RR-100, 1.7 µl sequencing buffer, and 3.2 pmol of forward or reverse PCR primer (see Note 15), made up to a total volume of 10 µl.

2. Reaction conditions are as follows: 25 cycles of rapid thermal ramp to 96 °C (1 °C/s), 96 °C for 30 s, rapid thermal ramp to 55 °C (1 °C/s), 55 °C for 20 s, rapid thermal ramp to 60 °C (1 °C/s), and 60 °C for 4 min.

3. Purify samples in sterile 96-well optical reaction plates with barcodes (Applied Biosystems, UK).

4. Precipitate DNA by the addition of 8 µl of H₂O followed by 32 µl ice-cold 95 % (v/v) ethanol.
<table>
<thead>
<tr>
<th>Target/enzyme</th>
<th>Expected PCR product (digestion product) band size (bp)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>TcI</td>
</tr>
<tr>
<td>LSU rDNA</td>
<td>110</td>
</tr>
<tr>
<td>HSP60/EcoRV</td>
<td>432−462 (432−462)</td>
</tr>
<tr>
<td>GPI/HhaI</td>
<td>1,264 (817+447)</td>
</tr>
</tbody>
</table>

$^a$According to Kawashita and others [77]
$^b$According to Brisse et al. [73]
$^c$For strains of North American origin, according to Brisse et al. [73]
$^d$Double band pattern observed for most isolates; 125 bp band exhibits variable intensity
5. Incubate samples at 4 °C for 15 min and then centrifuge for 45 min at 3,000 × g and 4 °C.
6. Remove the supernatant by inverting plates onto absorbent paper and centrifuging at 20 × g for 10 s.
7. Wash DNA pellets by the addition of 50 μl ice-cold 70 % (v/v) ethanol and briefly vortex.
8. Spin plates for 30 min at 3,000 × g and 4 °C.
9. Discard supernatants as previously.

10. Dry pellets at room temperature until no visible ethanol remains (see Note 16).

11. Resuspend DNA pellets in 10 μl Hi-Di™ deionized formamide (Applied Biosystems, UK) (see Note 14).

12. DNA pellets can be stored at −20 °C until required.

13. Analyze DNA sequences using an automated 16-capillary 3730 DNA Analyzer (Applied Biosystems, UK) (see Note 4).

3.5.5 Analysis of Nuclear MLST Data

Nucleotide data can be assembled manually in BioEdit v7.0.9.0 sequence alignment editor software (Ibis Biosciences, USA) [58] and ambiguous peripheral regions of aligned sequences discarded to produce unambiguous consensus sequences for each isolate. Heterozygous positions are identified by the presence of two coincident peaks at the same locus (“split peaks”), verified in forward and reverse sequences and scored according to the one-letter nomenclature for nucleotides from the International Union of Pure and Applied Chemistry (IUPAC). If data for multiple gene targets have been generated, sequences can be concatenated for each isolate (see Note 17). Distance-based phylogenies can be constructed using individual or concatenated heterozygous diploid sequence data in SplitsTree4 (select the average states parameter to handle ambiguous sites) [59]. To aid DTU assignment, a reference panel of sequences from Yeo et al. [20] is electronically available to download from GenBank. In the absence of a formalized nuclear MLST scheme for population genetic studies, three additional targets (LAP, RB19, and SODB), described in [20, 51, 52] can be used for higher resolution genetic diversity studies. Additional analyses are described with accompanying software by Tomasini et al. [80].

3.6 Maxicircle MLST

3.6.1 PCR Amplification

1. Prepare a 96-well PCR reaction plate (Fisher Scientific, UK) containing maxicircle primer stocks at 10 pmol/μl (see Table 3) arranged according to Fig. 2.

2. Amplify all ten maxicircle genes in standard PCR reactions each containing: 1× NH₄ reaction buffer, 1.5 mM MgCl₂ (Bioline, UK), 0.2 mM dNTPs (New England Biolabs, UK), and 1 U BIOTAQ™ DNA polymerase (Bioline, UK), made to a final volume of 17 μl.

3. Prepare a PCR mastermix for 90 samples without DNA template and aliquot 145 μl per well across the first plate row of a sterile 96-well PCR reaction plate (A01-A10).

4. Use a 10–100 μl twelve-channel pipette to transfer 17 μl mastermix per well down the 96-well PCR reaction plate (A01-H01, A02-H02, etc.).

5. Add 1 μl of DNA template (10–100 ng of T. cruzi genomic DNA) for each isolate across the plate (sample 1 in A01-A10, sample 2 in B01-B10, etc.).
6. Use a 0.5–10 μl twelve-channel pipette to transfer 1 μl of each forward and 1 μl of each reverse primer per well from the respective primer plates to the corresponding row on the PCR reaction plate (A01-A10, B01-B10, etc.) (see Notes 18 and 19).

7. PCR reactions are performed with an initial denaturation step of 3 min at 94 °C, followed by 30 amplification cycles (94 °C for 30 s, 50 °C for 30 s, 72 °C for 30 s) and a final elongation step at 72 °C for 10 min.

1. Visualize 5 μl of each PCR product by gel electrophoresis using 1.5 % agarose gels (Bioline, UK), as described in Subheading 3.4.2.

1. Purify all PCR products, as described in Subheading 3.4.3.

1. Use a 10–100 μl eight-channel pipette to transfer PCR products to a 96-well optical reaction plates with barcodes (Applied Biosystems, UK) for purification (see Note 19).

2. Sequence all PCR products, as described in Subheading 3.5.4.

Assemble sequence data as described for nuclear loci (see Subheading 3.5.5). For each isolate maxicircle sequences can be concatenated according to their structural arrangement (12S rRNA, 9S rRNA, CYT b, MURF1, ND1, COII, ND4, and ND5) and in the correct coding direction (see Note 17). The best-fit model of nucleotide substitution can be inferred in jMODELTEST 1.0 [60]. Phylogenies of increasing computational complexity can be constructed using MEGA 5 [61] (distance-based phylogenies), PhyML [62] (Maximum-Likelihood topologies) or MrBAYES v3.1 [63] (Bayesian topologies). A reference panel of maxicircle sequences is electronically available to download from GenBank under the accession numbers JQ581059-JQ581370 and JQ581403-JQ581480. For additional analyses please refer to Messenger et al. [56].

3.7 MLMT

3.7.1 PCR Amplification

1. Prepare a 96-well PCR reaction plate (Fisher Scientific, UK) with microsatellite primers diluted to 1 pmol/μl in 0.5× TE buffer (see Table 4) and arranged according to Fig. 3.

2. Amplify all microsatellite loci in a standard reaction containing: 1× ThermoPol Reaction Buffer (New England Biolabs, UK), 4 mM MgCl2, 34 μM dNTPs, 1 U Taq polymerase (New England Biolabs, UK), and 1 ng of genomic DNA, made up to a final volume of 7 μl.

3. Prepare one PCR mastermix (for 32 loci) per DNA isolate and aliquot 74 μl per well across A01-A04 of a sterile 96-well PCR reaction plate.
4. Each PCR plate can be used to amplify microsatellite loci for three DNA samples; distribute the mastermixes for isolates 2 and 3 across A05-A08 and A09-A12, respectively.

5. Use a 0.5–10 μl twelve-channel pipette to transfer 8.5 μl mastermix per well from A01-A12 down the PCR reaction plate (A01-H01, A02-H02, etc.)

6. Use a 0.5–10 μl twelve-channel pipette to transfer 1.5 μl of each premixed primer pair from the primer plate to the corresponding row on the PCR reaction plate (A01-A04, B01-B04, etc.).

7. Repeat step 6, instead transferring primers to columns 5–8 and 9–12.

8. PCR reactions for all loci are performed with an initial denaturation step of 4 min at 95 °C, then 30 amplification cycles (95 °C for 20 s, 57 °C for 20 s, 72 °C for 20 s) and a final elongation step at 72 °C for 20 min.

3.7.2 MLMT Multiplexing and Allele Size Determination

1. Use a 10–100 μl eight-channel pipette to combine columns 2, 3, and 4 into column 1, columns 6, 7, and 8 into column 5, and columns 10, 11, and 12 into column 9.

2. Transfer the contents of column 1, 5, and 9 into columns 1, 2, and 3 of a new sterile 96-well PCR reaction plate (see Fig. 6) to form a stock plate.

3. Each 96-well stock plate can hold multiplexed microsatellite PCR products from 12 DNA samples.

4. Mix 25 μl GeneScan™-500 LIZ™ fluorescent size standard (Applied Biosystems, UK) with 950 μl Hi-Di™ deionized formamide (Applied Biosystems, UK) and aliquot 82 μl per well across A01-A12 of a sterile 96-well optical reaction plate with barcode (Applied Biosystems, UK) (see Note 20).

5. Use a 0.5–10 μl twelve-channel pipette to distribute 9.75 μl of GeneScan™/Hi-Di™ solution into each well from A01-A12 down the 96-well optical reaction plate.

6. Use a 0.5–10 μl twelve-channel pipette to transfer 0.5 μl of sample PCR product from the stock plate into each corresponding row of the optical reaction plate (A01-A12, B01-B12, etc.).

7. Determine allele sizes using an automated 16-capillary sequencer (AB3730, Applied Biosystems, UK), with a standard injection time of 10 s.

3.7.3 Analysis of MLMT Data

Allele sizes can be assembled in GeneMapper® v 4.0 (Applied Biosystems, UK) and isolates should be typed “blind” to control for user bias and checked manually for errors. A set of allele sizes for reference strains and bin sizes for each microsatellite locus are available online at: http://www.ki.se/chagasepinet/mlmt.html
Fig. 6 Schematic of multiplexing microsatellite PCR products

Molecular Genotyping of *Trypanosoma cruzi* for Lineage Assignment and Population...
Microsatellite data are highly amenable to quantitative analysis. Population structures between different geographical areas and transmission cycles can be inferred using pair-wise distance-based measurements, such as $D_{AS}$ (infinite alleles model of IAM) or $\delta\mu^2$ (stepwise mutation model (SMM)) which can be calculated in MICROSAT v1.5d [64]. $D_{AS}$ values can be assembled into a distance matrix and used to construct Neighbor-Joining trees in PHYLIP v3.67 [65]. Support for nodes in the Neighbor-Joining tree can be generated in PHYLIP v3.67 using 1000 bootstrap replicates of the data generated in MICROSAT v1.5d. The mean number of alleles per locus (MNA) and the sum number of occurrences of specific alleles for each locus can be calculated using the Microsatellite Toolkit add-in [66] for MS Excel. The software FSTAT 2.9.3.2 [67] can be used to estimate sample-size corrected allelic richness ($A_r$) and the inbreeding coefficient $F_{is}$. Heterozygosity indices, including deviations from Hardy–Weinberg equilibrium, and the extent of population differentiation ($F_{ST}$) can be calculated in ARLEQUIN v3.0 [68]. Multilocus linkage disequilibrium, estimated by the Index of Association ($I_A$), can be calculated in MULTILOCUS v1.3b [69]. Mantel’s test to compare pair-wise geographical and genetic distances can be executed in GENALEX 6 [70].

We strongly discourage the use of model-based population assignment software (e.g., STRUCTURE and BAPS) as these programs use algorithms which assume Hardy–Weinberg expectations within populations and complete linkage equilibrium between genetic markers, two criteria that are largely violated by clonal reproduction in T. cruzi. Instead population subdivisions can also be inferred using a nonparametric (without Hardy–Weinberg constraints) $K$-means clustering algorithm [71], implemented in adegenet within the R 2.13 software package [78]. The number of “true” populations can be defined using the Bayesian Information Criterion (BIC) and the relationship between clusters can be evaluated using a Discriminant Analysis of Principal Components (DAPC), which first transforms allele frequencies at individual loci into uncorrelated variables (principal components), via a Principal Component Analysis (PCA) [72].

4 Notes

1. Infection with T. cruzi can only occur via direct inoculation or contamination of broken skin/intact mucosal membranes (conjunctiva, nose and mouth). Transmission via inhalation is highly unlikely as organisms do not readily aerosolize. In addition, parasites do not survive desiccation and are not free-living. Furthermore the predominant, but not exclusive form in exponentially growing axenic cultures is the non-infective epimastigote stage. To minimize risk of infection:
(a) Wear appropriate Personal Protective Equipment (PPE) at all times, including a Howie laboratory coat, eye-protection, and close-fitting disposable gloves.

(b) Conduct all manipulations of live material in a Class II microbiological safety cabinet, which should be fumigated regularly to prevent bacterial and fungal contamination.

(c) Do not touch the face or any exposed area while wearing contaminated gloves or handling live material.

(d) Routinely decontaminate work surfaces/cabinets with 70 % ethanol after use.

(e) Dispose of all contaminated material by immersing in 70 % ethanol or 10 % chloros (sodium hypochlorite) overnight.

(f) Restrict the use of sharps and glassware to avoid the risk of direct inoculation and dispose of all contaminated sharps in an appropriate sharpsafe bin.

(g) Avoid any procedures, e.g., centrifugation in open tubes or grinding of infected tissues, which may generate droplet suspensions.

(h) If necessary, wear a face visor or use a protective screen when directly handling infectious material, e.g., dissecting infected triatomine bugs.

(i) Establish full written risk assessments and emergency accident procedures before commencing work with live *T. cruzi*.

2. *T. cruzi* genomic DNA can be extracted from cultured epimastigotes, human hemocultures, or triatomine bug feces. The Gentra Puregene tissue kit (Qiagen, UK) and High Pure PCR template preparation kit (Roche, UK) both produce high quality template but with some loss of DNA yield and are most appropriate to extract DNA from cultured parasites and human clinical samples, respectively. DNA extracted using DNAzol® is typically of a higher yield but of lesser quality and is thus more suitable for extracting DNA from samples with low parasite density, including those derived from bug feces homogenate.

3. Ethidium bromide is mutagenic and toxic, so PPE must be worn at all times when handling this reagent.

4. We assume that the researcher has access to an automated fluorescent sequencer either through affiliations with an academic institution or by outsourcing to a commercial sequencing company.

5. Ideally, sterile test 5 % of each 4 N culture batch, by incubating at 37 °C for 3 days and checking for contamination.
6. Mercuric chloride is highly toxic and must be handled while wearing PPE (Howie laboratory coat, disposable gloves, and eye-protection) and with extreme care.

7. Some *T. cruzi* strains have a predilection to grow in clumps, therefore cultures should be checked microscopically and if clumpy, parasites can be separated by low-speed centrifugation (~200×*g*) prior to cloning.

8. Parasite genomic DNA can also be extracted from smaller culture volumes using the Gentra Puregene tissue kit. This protocol can be modified to extract DNA from 1 ml of *T. cruzi* culture in 1.5 ml graduated microcentrifuge tubes, by decreasing reagent volumes tenfold and performing all centrifugation steps in a microcentrifuge at >13,000×*g*.

9. Prepare NuSieve™ GTG™ low melting temperature agarose (Lonza, UK) by first soaking the agarose in chilled 1× TAE buffer for 15 min; this prevents the agarose from foaming during heating. Heat the agarose and buffer in a microwave on medium power for 2 min. Gently swirl the solution to resuspend any settled powder/gel pieces and reheat on high power until the solution begins to boil. Hold at boiling point for 1 min or until all of the agarose particles are dissolved. Allow the solution to cool to 50–60 °C prior to the addition of 0.5 μg/ml ethidium bromide and casting.

10. If consumable costs are restricted, PCR products can also be purified using absolute isopropanol. Add an equal volume of absolute isopropanol to PCR product in a sterile 0.5 ml graduated microtubes (Anachem, UK). Incubate at room temperature for 15 min. Spin tubes at >13,000×*g* in a microcentrifuge for 20 min and discard the supernatant. Wash the pellet in 70% (v/v) ethanol by spinning for 10 min. Discard the supernatant and air-dry the pellet. Resuspend the pellet in H₂O or 0.5× TE buffer.

11. Heating buffer EB to 55 °C before applying to the column and incubating for 1–5 min, prior to elution, can increase the yield from QIAquick columns.

12. If PCR-RFLP genotyping will be routinely performed it may be useful to prepare a stock of digested DNA size standards from *T. cruzi* reference isolates for each DTU. These can be stored at −20 °C and run as positive controls alongside unknown samples where necessary.

13. Ideally, *GPI*-RFLP gels should be run for long as possible in order to clearly separate bands at 490 and 447 bp (*TcV* and *TcVI* genotypes). In addition, the smallest *HSP60* band (118–148 bp; *TcIII*, *TcV*, and *TcVI* genotypes) can be difficult to
visualize, in which case it may be necessary to run a larger volume of digest reaction.

14. Aliquot both the Big Dye Sequencing RR-100 (e.g., 20 μl/ aliquot) and Hi-Di™ deionized formamide (e.g., 1 ml/aliquot) and store at −20 °C in order to minimize the number of freeze–thaw and exposure cycles for each tube. An appropriate volume aliquot will receive less than five freeze–thaw cycles and contain sufficient quantity for 1 week’s worth of reactions.

15. This modified protocol is for a reaction that is half the manufacturer’s recommended volume. In addition, the reagent mix (Big Dye Sequencing RR-100) has been reduced by one-eighth of the recommended amount to save considerable costs.

16. It is important to ensure that no ethanol remains in the sequencing reaction plate but equal care must be taken not to overdry the DNA pellets as this may inhibit their resuspension in Hi-Di™ deionized formamide.

17. Ensure that in the nucleotide alignment, isolate sequences are placed in the same order for each gene, otherwise it is very easy to mistakenly concatenate sequences from different isolates across multiple loci.

18. To speed manipulations, we strongly recommend the use of eight- and twelve-channel multichannel pipettes. Although it is possible to perform all pipetting individually, the multichannel renders the process much less laborious and more robust. Our current choice of pipette is the ErgoOne® range (Star Labs, UK) and we use 0.5–10 μl twelve-channel (S7112-0510), 10–100 μl twelve-channel (S7112-1100), and 10–100 μl eight-channel pipettes (S7108-1100) for maxicircle MLST and MLMT PCR amplifications.

19. When transferring between plates ensure that the plates are first lined up in the same orientation as each other (A01 to A01 and H12 to H12) as it is remarkably easy to accidentally reverse a plate.

20. It may be useful to create a set of allele size standards prepared from reference strains to run alongside samples as internal controls.

Acknowledgments

Research detailed in this chapter was funded by support from the Wellcome Trust, the BBSRC and the European Union Seventh Framework Programme grant 223034 (“ChagasEpiNet”).
References


64. Minch E, Ruiz-Linares A, Goldstein D et al (1997) MICROSAT v1.5d: a computer program for calculating various statistics on microsatellite allele data. Department of Genetics, Stanford University, Stanford, CA


68. Excoffier L, Schneider S (2005) Arlequin ver. 3.0: an integrated software package for population genetics data analysis. Evol Bioinform Online 1:47–50


77. Kawashita SY, Sanson GFO, Fernandes O et al (2001) Maximum-likelihood divergence date estimates based on rRNA gene sequences sug-
gests two scenarios of *Trypanosoma cruzi* intraspecific evolution. Mol Biol Evol 18: 2250–2259


COVER SHEET FOR EACH ‘RESEARCH PAPER’ INCLUDED IN A RESEARCH THESIS

Please be aware that one cover sheet must be completed for each ‘Research Paper’ included in a thesis.

1. For a ‘research paper’ already published

1.1. Where was the work published? BMC Genomics

1.2. When was the work published? 2012

1.2.1. If the work was published prior to registration for your research degree, give a brief rationale for its inclusion

...........................

...........................

1.3. Was the work subject to academic peer review? Yes

1.4. Have you retained the copyright for the work? Yes

If yes, please attach evidence of retention.

If no, or if the work is being included in its published format, please attach evidence of permission from copyright holder (publisher or other author) to include work.

2. For a ‘research paper’ prepared for publication but not yet published

2.1. Where is the work intended to be published?

2.2. Please list the paper’s authors in the intended authorship order

2.3. Stage of publication – Not yet submitted / Submitted / Undergoing revision from peer reviewers’ comments / In press

3. For multi-authored work, give full details of your role in the research included in the paper and in the preparation of the paper. (Attach a further sheet if necessary)

The candidate was responsible for assembling the whole maxicircle genome sequence for T. c. marinkellei and performing all associated mitochondrial analyses. The candidate drafted all sections pertaining to mitochondrial data for the final manuscript.

NAME IN FULL (Block Capitals) LOUISA ALEXANDRA MESSENGER

STUDENT ID NO: 223021

CANDIDATE’S SIGNATURE Date 10/2/15

SUPERVISOR/SENIOR AUTHOR’S SIGNATURE (3 above)
Comparative genomic analysis of human infective Trypanosoma cruzi lineages with the bat-restricted subspecies T. cruzi marinkellei

Oscar Franzén1*, Carlos Talavera-López1, Stephen Ochaya1, Claire E Butler2, Louisa A Messenger3, Michael D Lewis3, Martin S Llewellyn3, Cornelis J Marinkelle4, Kevin M Tyler2, Michael A Miles3 and Björn Andersson1*

Abstract

Background: Trypanosoma cruzi marinkellei is a bat-associated parasite of the subgenus Schizotrypanum and it is regarded as a T. cruzi subspecies. Here we report a draft genome sequence of T. c. marinkellei and comparison with T. c. cruzi. Our aims were to identify unique sequences and genomic features, which may relate to their distinct niches.

Results: The T. c. marinkellei genome was found to be ~11% smaller than that of the human-derived parasite T. c. cruzi Sylvio X10. The genome size difference was attributed to copy number variation of coding and non-coding sequences. The sequence divergence in coding regions was ~7.5% between T. c. marinkellei and T. c. cruzi Sylvio X10. A unique acetyltransferase gene was identified in T. c. marinkellei, representing an example of a horizontal gene transfer from eukaryote to eukaryote. Six of eight examined gene families were expanded in T. c. cruzi Sylvio X10. The DGF gene family was expanded in T. c. marinkellei. T. c. cruzi Sylvio X10 contained ~1.5 fold more sequences related to VIPER and L1Tc elements. Experimental infections of mammalian cell lines indicated that T. c. marinkellei has the capacity to invade non-bat cells and undergo intracellular replication.

Conclusions: Several unique sequences were identified in the comparison, including a potential subspecies-specific gene acquisition in T. c. marinkellei. The identified differences reflect the distinct evolutionary trajectories of these parasites and represent targets for functional investigation.

Background

The subgenus Schizotrypanum harbors the type species Trypanosoma cruzi, which is the causative agent of Chagas disease in humans. Other members of the Schizotrypanum subgenus are often referred to as T. cruzi-like species as they are morphologically similar or indistinguishable from T. cruzi [1]. With the exception of the human infecting parasite, members of Schizotrypanum are restricted to bats (order Chiroptera) and occur in high prevalence among bats in Latin America and elsewhere in the world [1-4]. There is no evidence that T. cruzi-like parasites are harmful to bats, although this may reflect a paucity of data. Most infected bats are insectivorous and infection is thought to take place either through ingestion of infected arthropods or via stercorarian transmission from bat-feeding bugs [5,6]. The genetic diversity of T. cruzi-like species and their evolutionary relationships are yet to be determined.

Trypanosoma cruzi marinkellei is a bat-associated subspecies of T. cruzi [1]. The human infective parasite T. cruzi should accordingly be referred to as the nominate subspecies T. cruzi cruzi (T. c. cruzi) [1]. T. c. marinkellei is prevalent among bats in Central and South America, which are its only known mammalian hosts [1,5]. It differs from T. c. cruzi in terms of isoenzyme electrophoresis patterns and buoyant DNA densities. T. c. marinkellei does not infect immunocompetent mice [1,5], nor does it provide immunological protection against challenge with T. c. cruzi [1], suggesting that the infection is characterized by distinct antigenic profiles. Sequence-based phylogenies have confirmed a relatively close relationship with T. c. cruzi [5,7-9] and estimated the divergence time at
~6.5-8.5 MYA [10-12]. Cavazzana et al. reported that *T. c. marinkellei* was associated with phyllostomid species (insectivorous, frugivorous, carnivorous and haematophagous bats) [5] and transmission is thought to occur when triatomine bugs of the genus *Cavernicola* feed on bats [13]. However, the natural transmission cycle among bats is not well characterized and there might be other vectors or direct transmission mechanisms. Some genetic structure within the *T. c. marinkellei* population has been reported [14], but the strength of correlation between parasite lineage and host remains to be defined. Moreover, bat-restricted parasites are of evolutionary interest, since it has been proposed that *T. c. cruzi* may have originated from an ancestral bat-lineage that jumped into terrestrial mammals [15]. The present day human lineage, *T. c. cruzi*, has been in contact with humans for no more than 10,000 to 30,000 years, which is the period of human presence in the Americas [16].

*T. c. cruzi* strains are currently sorted into six lineages or discrete typing units (DTUs), which illustrate the genetic diversity of this parasite [17]. Several strains have to date been subjected to genome sequencing, among these are CL Brener and Sylvio X10. The CL Brener strain was selected for the original genome project and belongs to DTU VI. The size of the CL Brener genome was ~110 Mb and it was assembled mostly with Sanger paired-end reads. The CL Brener strain was shown to be a genetic hybrid of two diverged haplotypes named Esmeraldo-like and non-Esmeraldo-like [18]. The hybrid and repetitive nature of this genome complicated sequence assembly and finishing, leaving the genome in many gapped scaffolds and contigs. Weatherly et al. later compiled scaffolds into more complete chromosome-wide sequences [19]. Second-generation sequencing facilitates more cost-effective and rapid sequencing efforts. Recently, 454 sequencing was applied on the genome of the DTU I strain Sylvio X10 [20], revealing a slightly smaller but still repeat-rich genome.

Little is known about genomic variation among organisms within the *Schizotrypanum* genus. Genomic insights can provide information on evolutionary adaptation of these parasites, as well as being useful for advancing population genetics. Thus, exploring genomic diversity could reveal important genetic and biological characteristics, and potentially clues as to how these parasites relate to the human disease. Here we describe the genome of *T. c. marinkellei* B7, a bat-associated parasite originally isolated from a colony of the pale spear-nosed bat *Phyllostomus discolor* roosting in a hollow tree [1]. The parasite was isolated in São Felipe, Bahia state, Brazil in 1974 and has since then been stored under cryogenic conditions with occasional short periods of *in vitro* cultivation. We combined Roche/454 and Illumina sequencing to generate a draft genome sequence of *T. c. marinkellei*. This is the first whole genome analysis of a *T. c. cruzi*-like species that is not associated with human infections. In addition, we also report re-assembly and re-annotation of the human infective strain *T. c. cruzi* Sylvio X10 [20], a commonly used reference strain of *T. c. cruzi* I [21], using additional sequence data.

The comparative analyses with *T. c. cruzi* revealed that the genomes contain the same repertoire of housekeeping genes. Moreover, *T. c. marinkellei* contains an additional gene that appears to be an example of recent horizontal gene transfer. In addition, the genomes also exhibit copy number variation and diversification of gene families, which potentially give rise to a large number of strain-specific protein isoforms.

### Results and discussion

#### Sequencing and Assembly of *T. c. marinkellei* and *T. c. cruzi* Sylvio X10

In the text, we refer to *Trypanosoma cruzi marinkellei* as Tcm, *Trypanosoma cruzi* Sylvio X10 as Tcc X10 and *Trypanosoma cruzi* cruzi CL Brener as Tcc CLBR. Genomic sequence reads were generated from Tcm and Tcc X10 using 454 and Illumina sequencing (Table 1). 454 sequencing (single end; long reads) was performed on genomic DNA from Tcm, which produced sequence reads with an average length of ~357 nt. The 454 data from Tcc X10 was the same as previously described [20]. In addition, one ~2 kb insert library (2×100-nt reads) was prepared for Tcm and Tcc X10 respectively, using a modified version of the Illumina mate-pair protocol (Materials and Methods). The modified Illumina protocol was chosen to enable 100 nt read lengths, as Illumina does not recommend its own protocol for mate-pair sequencing with read lengths >36 nt. This generated 71,948,029 and 84,638,048 read-pairs from Tcm and Tcc X10 respectively. Not all read-pairs translated to the expected insert size of ~2 kb. Long insert libraries often contain a significant proportion of short insert fragments (corresponding to paired-end reads). Most often this is due to non-optimal biotin enrichment causing some fragments not to circularize and therefore become sequenced with much shorter insert. We determined the number of

<table>
<thead>
<tr>
<th>Table 1 Raw sequence data</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td><strong># reads (10⁶)</strong></td>
</tr>
<tr>
<td><strong># nt (10⁹)</strong></td>
</tr>
<tr>
<td><strong>Average read length (nt)</strong></td>
</tr>
<tr>
<td><strong>Coverage</strong></td>
</tr>
</tbody>
</table>

*Single end 454 reads,* b No. Read-pairs (true mate-paired reads after adapter trimming), c Billion nucleotides.* The average read length (after adapter trimming). *The theoretical genome coverage based on known genome sizes and the number of sequenced nucleotides.*
true mate-pairs from the obtained data using an R-script previously published by Van Nieuwerburgh et al. [22]. The script determines the location of the LoxP linker sequence in the read, and then uses this information to classify read-pairs as true mate-pairs, paired-end, single-end or linker-negative. True mate-pairs should contain the LoxP sequence close to the 3’ end in at least one read, indicating that circularization has taken place. In our data, 32% (23,055,208/71,948,029) and 34% (28,781,049/84,638,048) of the read-pairs were classified as true mate-pairs from Tcm and Tcc X10 respectively (LoxP sequence close to the 3’ end in at least one of the reads). 38% (27,890,116/71,948,029) and 35% (30,076,419/84,638,048) read-pairs were classified as paired-end from Tcm and Tcc X10 respectively. The remaining read-pairs were either unpaired or LoxP-negative, meaning that the linker was present in the unsequenced part of the fragment or that the fragment did not contain a linker. Hence, despite an improved protocol, a substantial number of paired-end and single-end reads were obtained. The causes of this has previously been discussed [22].

The 454 and Illumina data were subsequently assembled (Figure 1). In order to take platform dependent sequencing artifacts into consideration, 454 and Illumina reads were assembled separately using different assembly programs (Figure 1; Table 2; Materials and Methods). Insertion-deletion errors in the 454 assemblies were identified and corrected using alignments with Illumina reads, which corrected 12,358 and 7,277 positions of Tcm and Tcc X10 respectively. The most common error was one or two missing bases (~90% of the corrected positions). The resulting assemblies were subsequently merged into a non-redundant assembly. Distance information from mate-pair reads was used to arrange contigs into scaffolds. Where possible, the distance between two adjacent contigs in a scaffold was inferred by comparison with Tcc CLBR, i.e. if two contigs flanking each side of a gap could be aligned with one of the CL Brener haplotypes, then the approximate gap length could be inferred from CL Brener. As a final assembly step, both Tcm and Tcc X10 were subjected to gap closure using the IMAGE pipeline [23] and the sorted paired-end reads (see above). Prior to feeding scaffolds into IMAGE, paired-end reads were quality filtered. IMAGE uses iterative mapping of reads to contig ends, followed by local assembly and alignment to close gaps and extend contigs. Eight IMAGE iterations were completed for each genome, which improved each assembly by adding 653,655 (Tcm) and 534,614 (Tcc X10) base pairs, which closed 261 and 171 gaps and extended 2,426 and 2,510 contig ends from Tcm and Tcc X10 respectively.

The combined assembly lengths were 38.6 Mb and 43.4 Mb for Tcm and Tcc X10 respectively. The assembly size of Tcc X10 was very similar to our previous estimate from extrapolation of unassembled data [20] and flow cytometry [24]. Flow cytometry analysis estimated the haploid size of Tcm to ~39 Mb (Additional file 1: Figure S1), which was close to the in silico assembly length. Thus, assembly sizes were consistent with experimental measurements. Moreover, this confirmed that the Tcm genome was ~4.8 Mb smaller than that of Tcc X10. The percentage of assembled bases in each assembly was very similar: Tcm 88.6% (34.2 Mb/38.6 Mb); Tcc X10 88.7% (38.5 Mb/43.4 Mb). We analyzed 29,422 unused 454 reads of Tcm with RepeatMasker, which identified 13,108 reads corresponding to kinetoplastid sequences. The remaining reads were analyzed with BLAST, showing them to correspond multicopy genes or other repeats.

868 (Tcm) and 987 (Tcc X10) scaffolds were longer than 5 kb, which corresponded to 25.7 and 26.8 Mb (including gaps). The longest scaffolds were 335 kb (Tcm) and 384 kb (Tcc X10). Some 200 gaps could be closed from the apparent overlap of adjacent contigs. Compared with 454 reads alone, addition of mated reads provided longer contigs and scaffolds, corrected 454 sequence errors and allowed accurate estimation of genome heterozygosity and copy number variation.

**Figure 1 Schematic overview of the sequence assembly.**

Schematic overview of the genome assembly steps. Illumina reads were assembled into contigs with Velvet. Unused reads were extracted and used for a second Velvet assembly with a different kmer length. 454 reads were assembled with CELERA. The 454-assemble was then subjected to homopolymer error correction with Illumina reads. The Illumina and 454 assemblies were merged into a non-redundant assembly using the Zorro pipeline. The assembly was then subjected to scaffolding using SSPACE and physical distance information. The final step involved gap closure with the IMAGE pipeline.

**Comparison of heterozygosity and multicopy genes**

The level of heterozygosity among populations of medically important trypanosomes is likely to reflect the impact of key evolutionary processes such as gene conversion and genetic exchange. In the present study we estimated the amount of heterozygosity in Tcm and Tcc
Table 2 Genome assembly statistics and summary

<table>
<thead>
<tr>
<th>Genome</th>
<th>Step</th>
<th>Software</th>
<th>Size</th>
<th># contigs</th>
<th># scaffolds</th>
<th>Average length</th>
<th>N50</th>
<th>N90</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tcm</td>
<td>454 assembly</td>
<td>CELERA</td>
<td>37.3</td>
<td>30,737</td>
<td>-</td>
<td>1,216</td>
<td>1,670</td>
<td>539</td>
</tr>
<tr>
<td></td>
<td>Illumina assembly</td>
<td>Velvet (kmer 43)</td>
<td>16.7</td>
<td>9,247</td>
<td>-</td>
<td>1,813</td>
<td>2,378</td>
<td>851</td>
</tr>
<tr>
<td></td>
<td>Assembly of non-assembled Illumina reads</td>
<td>Velvet (kmer 53)</td>
<td>1.17</td>
<td>2,094</td>
<td>-</td>
<td>562</td>
<td>536</td>
<td>418</td>
</tr>
<tr>
<td></td>
<td>Assembly merging</td>
<td>Zorro</td>
<td>33.5</td>
<td>24,799</td>
<td>-</td>
<td>1,353</td>
<td>2,218</td>
<td>549</td>
</tr>
<tr>
<td></td>
<td>Scaffolding</td>
<td>SSPACE</td>
<td>38.8</td>
<td>23,813</td>
<td>1,835</td>
<td>2,296</td>
<td>25,044</td>
<td>576</td>
</tr>
<tr>
<td></td>
<td>Gap closure</td>
<td>IMAGE</td>
<td>38.6</td>
<td>23,000</td>
<td>1,774</td>
<td>2,302</td>
<td>25,781</td>
<td>583</td>
</tr>
<tr>
<td>Tcc X10</td>
<td>454 assembly</td>
<td>CELERA</td>
<td>41.8</td>
<td>33,686</td>
<td>-</td>
<td>1,243</td>
<td>1,516</td>
<td>549</td>
</tr>
<tr>
<td></td>
<td>Illumina assembly</td>
<td>Velvet (kmer 43)</td>
<td>17.0</td>
<td>8,523</td>
<td>-</td>
<td>1,997</td>
<td>2,742</td>
<td>904</td>
</tr>
<tr>
<td></td>
<td>Assembly of non-assembled Illumina reads</td>
<td>Velvet (kmer 53)</td>
<td>1.14</td>
<td>2,116</td>
<td>-</td>
<td>543</td>
<td>523</td>
<td>416</td>
</tr>
<tr>
<td></td>
<td>Assembly merging</td>
<td>Zorro</td>
<td>38.0</td>
<td>28,389</td>
<td>-</td>
<td>1,339</td>
<td>1,869</td>
<td>560</td>
</tr>
<tr>
<td></td>
<td>Scaffolding</td>
<td>SSPACE</td>
<td>43.7</td>
<td>27,605</td>
<td>2,476</td>
<td>2,162</td>
<td>14,067</td>
<td>589</td>
</tr>
<tr>
<td></td>
<td>Gap closure</td>
<td>IMAGE</td>
<td>43.4</td>
<td>26,889</td>
<td>2,423</td>
<td>2,158</td>
<td>14,516</td>
<td>592</td>
</tr>
</tbody>
</table>

* The length when sequences are combined (Mb).
* The number of contigs/scaffolds.
* The average contig length (bp). For the SSPACE row, this refers to the average scaffold length.
* The length N for which half of all bases are in a sequence of this length or longer.
* The length N for which 90% of all bases are in a sequence of this length or longer.
* Contigs >500 bp.

X10 by aligning Illumina and 454 reads back to the assemblies and subsequently identifying high quality mismatches between the consensus sequence and aligned reads. In order to increase the confidence, only nucleotide positions with 10 to 80X coverage were included and contigs shorter than 5 kb were ignored. This resulted in 19,015,919 and 20,468,447 positions of Tcm and Tcc X10 that permitted analysis, which represented 38.8% (14,712 bp/37,894 bp), of which 38.8% (14,712 bp/37,894 bp) and 42.4% (19,513 bp/46,001 bp) were located in protein-coding genes. 7,976 and 10,596 heterozygous positions of Tcm and Tcc X10 were located at non-synonymous sites. Gene Ontology analysis was performed on genes containing at least one polymorphism at a non-synonymous site, resulting in two significantly enriched categories (p<0.05): GO:0009451 (RNA modification) and GO:0009982 (pseudouridine synthase activity). Overall, the estimated level of heterozygosity of Tcc X10 was slightly higher than previously reported [20], likely due to the increased sequence depth in the present study. In order to identify regions with higher density of heterozygosity, i.e. clustering of heterozygous sites, we counted the number of heterozygous positions inside 1,000 bp windows. This indicated that heterozygosity often, but not exclusively, was located in clusters (Figure 2). In conclusion, heterozygosity of the Tcm and Tcc X10 were ~0.19% and ~0.22%, with some regions exhibiting higher than average heterozygosity. In contrast, the heterozygosity level of Tcc CLBR was ~1 to 4% (since it is a hybrid). In comparison to other kinetoplastids, the heterozygosity level is similar to that of Leishmania braziliensis but higher than L. major and L. infantum [25]. The generally low levels of heterozygosity found in many protozoans is difficult to explain in terms of a strictly clonal propagation model [10]. Such organisms would be expected to observe extensive divergence of homologous genomic copies, which is the case for bdelloid rotifers [26]. In perspective, the B lineage of the human parasite Giardia intestinalis exhibits relatively...
high heterozygosity (~0.5%) [27] whereas A and E lineages exhibit low heterozygosity (~0.01%) [28]. The genome of the free-living amoebflagellate Naegleria gruberi was described as mosaic of homozygous and heterozygous regions, with an average polymorphism rate of 0.58% [29]. Interestingly, asexual lineages of Daphnia exhibit low levels of allelic divergence and appear to employ ameiotic recombination to eliminate heterozygosity faster than it accumulates [30]. The mechanism for maintaining low heterozygosity in trypanosomatids remains unknown, but could involve cryptic sexuality, frequent local gene conversion or chromosome-wide conversion. The former can be evaluated via an assessment of population-level inter-locus linkage disequilibrium. Nevertheless, descriptive data may not be sufficient to explain the causes of this phenomenon.

T. c. cruzi contain several highly expanded and complex gene families [31,32], comprised of transcribed genes and pseudogenes. Several of these families have been reported to vary in copy number between strains [18,20,33-35]. In the present study we compared gene family content between Tcm and Tcc X10 using the depth of aligned short reads. Initially, repeat boundaries were determined using RepeatMasker. Subsequently, the percentages of reads mapping to repeat families were calculated (Table 3). The statistical significance was assessed in the following way: an empirical distribution of genome-wide read-depth differences was established using regions of homology between Tcm and Tcc X10 (Additional file 2: Figure S2). The software promer was used to find homologous regions. In each homologous region, the percentage read coverage was calculated for Tcm and Tcc X10. These numbers were then corrected for the genome size and the coverage difference for each homologous region was computed. 17,290 regions were included, with mean 1.380393e-07 and standard deviation 5.83481e-07. The logspline function of the R package with the same name was used to fit a smoothed density curve to the data, and the plogspline function was used to determine p-values. Six out of eight examined gene families were expanded in Tcc X10: trans-sialidase; mucin-associated surface protein; retrotransposon hot spot protein; TcMUC mucin; ABC Transporter; and RNA binding protein. On the contrary, GP63 and dispersed gene family 1 (DGF) were contracted in Tcc X10. The observation that DGF was contracted in Tcc X10 was consistent with previous data when Tcc X10 was compared with Tcc CLBR [20], which suggests a recent loss of DGF-related sequences in the lineage leading to Tcc X10. Further examination of several DTU I strains may resolve if this is a general feature of this lineage. It is clear that at least part of the genome size difference can be attributed to expansion and/or contraction of these gene families. We performed a closer examination of the TcMUCII mucin gene family. TcMUCII mucin genes of the same genome were frequently found to be too different to align. We constructed entropy plots from alignment positions that were deemed as accurate, which revealed, as expected that 5' and 3' termini were more conserved and the internal parts of these genes were hypervariable (Additional file 3: Figure S3).

**Kinetoplastid DNA (maxicircle)**

The mitochondrial genomes (maxicircles) of T. c. cruzi strains X10 (DTU I), Esmeraldo (DTU II) and CLBR (DTU VI) have been sequenced, and have provided insights into the structure and organization of kinetoplastid DNA of these strains [36]. The T. c. marinkellei maxicircle

<table>
<thead>
<tr>
<th>Table 3 Comparison of gene family content</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>T. c. marinkellei</strong></td>
</tr>
<tr>
<td>Gene family a</td>
</tr>
<tr>
<td>DGF</td>
</tr>
<tr>
<td>TS</td>
</tr>
<tr>
<td>MASP</td>
</tr>
<tr>
<td>RHS</td>
</tr>
<tr>
<td>GP63</td>
</tr>
<tr>
<td>TcMUC mucin</td>
</tr>
<tr>
<td>ABC</td>
</tr>
<tr>
<td>RBP</td>
</tr>
</tbody>
</table>

- a The combined number of base pairs of this gene family that was identified in the assembly. Sequences were identified using RepeatMasker and a repeat library of coding sequences from the Tcc CLBR genome. These numbers include partial coding sequences. The number inside parenthesis refers to the percentage of total assembly size.
- b The percentage of short reads that mapped to these features.
- c SE=Significantly Enriched. Refers to if one genome contained significantly more of this gene family. The significance was determined from an empirical distribution of read depth differences from homologous regions of Tcm and Tcc X10, corrected for genome size. The empirical distribution was used to calculate a p-value.
was identified as a 20,037 bp contig from the 454 assembly. The length of this sequence was slightly longer (~ 5 kb) than those previously reported, and the difference was attributed to variability in the repetitive region. The coding region of the Tem maxicircle was syntenic with the coding regions of the three complete T. c. cruzi maxicircle genomes, beginning with the 12S rRNA gene and ending with the ND5 gene. The lengths of the individual genes within the Tem maxicircle coding region were comparable to those of the three T. c. cruzi strains (Additional file 4: Table S1). The length of the complete maxicircle coding region (beginning at 12S rRNA and ending after ND5) for Tem was 15,438 bp and began after 4,599 bp of non-coding sequence. With respect to coding sequences, the average maxicircle nucleotide identity between Tem and Tcc X10 was (mean ± sd): 85.12% ± 6.1, between Tem and Tcc CLBR was 85.4% ± 6.2 and between Tem and Tcc Esmeraldo was 85.3% ± 6.1 (Additional file 4: Table S1). Phylogenetic reconstruction of the maxicircles from Tem, Tcc X10, Tcc CLBR and Tcc Esmeraldo confirmed that the Tem maxicircle was slightly closer to Tcc Esmeraldo than Tcc X10/CLBR (Additional file 5: Figure S4). The topology of the tree suggests that the Esmeraldo maxicircle might represent the ancestral maxicircle lineage of T. c. cruzi.

The consensus maxicircle genome sequence is derived from the predominant nucleotide present across multiple read alignments at each position. However, this criterion disregards low abundance single nucleotide polymorphisms (SNPs) and therefore masks minor maxicircle haplotypes (heteroplasmy), which has previously been reported from Tcc X10 [37]. Illumina reads were used to assess the presence/absence of minor Tem maxicircle haplotypes. In total, this identified 19,821 reads that aligned to the Tem maxicircle. Low levels of heteroplasmy were observed in the Tem maxicircle protein-coding region. Twenty SNPs were identified among four genes (ND8, MURF1, COI and ND3) and one intergenic region (between CR4 and ND4). Average read depth for each SNP site was 47. At heterozygous sites, the minor nucleotide was present among an average of 9.5% (± 3.3%) of reads. All SNPs were bi-variable except for at two intergenic positions, where two minor nucleotides were present. These observations imply the occurrence of at least two minor mitochondrial haplotypes.

Gene content analysis and comparison

The Tem and Tcc X10 genomes were annotated using a semi-automatic strategy, which relied on the previous annotation of the reference genome Tcc CLBR [18]. Gene models were transferred from Tcc CLBR to Tem and Tcc X10 using Perl scripts, reciprocal BLASTp searches together with positional information (Materials and Methods). In addition, gene prediction was performed and gene models were kept if one or more of the following criteria were satisfied: (i) the gene was conserved in a syntenic position in Tcc CLBR; (ii) the gene shared homology with one or more gene families in Tcc CLBR; and (iii) the gene was longer than 250 amino acids. Gene models with complete overlap with another gene were discarded. The final annotations were manually inspected and refined with the Artemis Comparison Tool [38]. After this procedure, the genome sequences contained 10,342 (Tem) and 11,112 (Tcc X10) protein coding gene annotations, of which 60.5% (6,267/10,342) and 57.7% (6,416/11,112) were syntenic with Tcc CLBR, Tem and Tcc X10 respectively. With respect to coding sequences, the average nucleotide identity between Tem and Tcc X10 was 92.5% ± 3.2 (Figure 3). When Tem was compared with Tcc CLBR Esm and non-Esm the average nucleotide identity was 92.8% ± 3.4 and 92.6% ± 3.2. These identities are based on a total of 6,283 (Tem:Tcc X10), 5,441 (Tem: Tcc CLBR Esm) and 5,617 (Tem:Tcc CLBR non-Esm) orthologous gene pairs. ~86% (14,997/17,332) of the analyzed ortholog pairs had a nucleotide identity of 90% or higher. The ratio of non-synonymous and synonymous nucleotide substitutions ($\omega = dN/dS$) was 0.31 ± 0.21 in
average (Tcm versus Tcc CLBR non-Esm), indicating as expected that most genes were under purifying (stabilizing) selection. A total of 69 genes showed ω values > 1.1, indicating positive selection (Additional file 6: Table S2).

In order to identify isolate-specific genes, we compared the gene complements of the three genomes (Tcm, Tcc X10 and Tcc CLBR) using BLAST. Initially, the predicted proteomes of Tcm and Tcc X10 were queried with BLASTp against the predicted proteome of Tcc CLBR using the E-value threshold 1e-20. This resulted in 237 (Tcm) and 290 (Tcc X10) proteins longer than 250 amino acids that were not found in the Tcc CLBR proteome. These protein sequences were queried using tBLASTn against the Tcc CLBR genome to exclude the possibility that these putative genes were present as non-annotated open reading frames, using the same E-value threshold. This decreased the number of hits to 22 (Tcm) and 3 (Tcc X10). The composition of the 22 putative Tcm-specific genes were as follows: 11 TcMUCII mucin genes; 1 acetyltransferase (MOQ_006101); 5 putative genes with weak hits to microbial sequences (MOQ_006053, MOQ_007485, MOQ_009774, MOQ_006631, MOQ_003304); and 5 putative genes with no hits in public databases (MOQ_003636, MOQ_005225). For Tcc X10, one of the specific genes corresponded to a diverged mucin-associated surface protein (TCSYLVIO_008353). The remaining two putative genes did not show any significant hits in public databases (TCSYLVIO_011068 and TCSYLVIO_008789). Thus, the improved Tcc X10 genome sequence facilitated the detection of two putative Tcc X10-specific protein-coding genes not apparent in the earlier version. The two unknown genes were found to also be present in the previously reported draft genome sequence of Tcc X10 [20].

We used the same strategy to perform the reversed search, i.e. searching for genes specific for Tcc CLBR. This resulted in 344 and 206 protein sequences that were not found in Tcm and Tcc X10. Searches using tBLASTn towards Tcm and Tcc X10 further decreased this number to 70 and 100, and of these 52 and 21 were mucin-associated surface proteins or TcMUCII mucin. 8 (Tcm) and 26 (Tcc X10) contained low complexity repeats. The remaining 10 (Tcm) and 53 (Tcc X10) genes were queried against the raw 454 reads of Tcm and Tcc X10, which further decreased the number of Tcc CLBR specific genes to 3 that were not present in Tcm (Tc00.1047053511585.110, Tc00.1047053509525.260, Tc00.1047053510073.24). The 3 genes were uncharacterized (hypothetical). The Tcc CLBR-specific genes, compared with Tcc X10, were identical to those previously reported [20]. In conclusion, the total number of specific genes was remarkably low in relation to the number of coding sequences in these genomes. As a perspective, comparative genomics of T. brucei brucei and T. brucei gambiense did not identify any gene that could explain the ability to infect different species, despite interspecific pathological variation [39].

A specific acetyltransferase gene in T. c. marinkellei

As mentioned above, a 1,662 bp acetyltransferase gene (MOQ_006101) was found among the 22 unique genes in Tcm. This gene was identified in a single copy on scaffold 2842 and was missing in T. c. cruzi. Alignment of scaffold 2842 from Tcm with Tcc CLBR showed that it aligned close to the end of chromosome 37 and was flanked by VIPER elements and an ATPase gene (Figure 4). To exclude the possibility that MOQ_006101 was not properly assembled in T. c. cruzi, we searched raw 454/Illumina reads from Tcc X10 and raw Sanger

Figure 4 Genomic location of the T. c. marinkellei specific acetyltransferase gene (MOQ_006101). Screenshot from Artemis Comparison Tool of a ~19 kb homologous region of T. c. marinkellei B7 (contig 2842) and T. c. cruzi CL Brener (non-Esmeraldo-like haplotype; chr 37). Vertical black lines in each frame represent stop codons. Genes with shared homology in both genomes are shown in blue and the specific T. c. marinkellei gene (MOQ_006101) is shown in green. Red stripes represent regions with high sequence similarity between the two genomes. Abbreviations: ETIF (eukaryotic translation initiation factor 3 subunit 8, putative); HP (hypothetical protein); PDI (protein disulfide isomerase); CO (cytochrome c oxidase subunit IX); AT (acetyltransferase); and RP (U1A small nuclear ribonucleoprotein).
reads from Tcc CLBR. This confirmed that MOQ_006101 was not present in these genomes. Domain searches of MOQ_006101 revealed the presence of a Cas1p domain (pfam07779, E-value=9e-66) and multiple trans-membrane domains. In GenBank, the best hit from protein BLAST was to the green algae Chlamydomonas reinhardtii, containing 44% sequence identity over 496 amino acids (E-value < 4e-125). 4 iterations of PSI-BLAST resulted in hits to various species of plants and algae. The best ten hits were to the enzyme O-acetyltransferase, displaying protein identities between ~37-39% (Table 4). This indicated that MOQ_006101 has either diverged since the transfer to Tcm or that it has been transferred from a species not contained in GenBank, of which the latter seems the most likely. Furthermore, transcription of MOQ_006101 was detected with reverse transcriptase quantitative polymerase chain reaction (RT-qPCR).

Phylogenetic reconstruction of MOQ_006101 demonstrated that the closest known homologs were from various species of algae and plants (Figure 5A), and the absence of exon-intron boundaries suggested that it was transferred as a spliced mRNA. No homologs were found in Trypanosoma rangeli (Edmundo C. Grisard, Personal communication), Rhodnius prolixus (insect vector) or Myotis lucifugus (a bat species). The GC content of MOQ_006101 was 42.8%, which was significantly lower than the average of 52.7% ± 5.8 (Figure 5B). The GC content of the first, second and third codon positions were 42.2%, 37.7% and 48.6%, in contrast to the global levels: 57.5% ± 5.4 (GC1), 45.0% ± 6.2 (GC2) and 55.7% ± 11.0 (GC3). Hence, the GC content of MOQ_006101 was unusually low in relation to the global GC content of all Tcm genes. In conclusion, this suggested that the nucleotide composition of MOQ_006101 was distinct compared with all other genes of the Tcm genome. The unusual GC content can be interpreted as an imprint from the originating genome.

Codon Adaptation Index (CAI) is a measure of synonymous codon usage bias and can be used to evaluate the extent of which codon usage of a supposed foreign gene is similar to highly expressed genes in the host genome [40]. CAI can range between 0 and 1 and values closer to 0 imply equal use of synonymous codons whereas values closer to 1 imply strong codon usage bias. MOQ_006101 displayed a CAI value of 0.518 (Figure 5C). In contrast, the median CAI across all genes was 0.545 ± 0.05 (median ± median absolute deviation). Thus, CAI was lower than the mean but still within the expected range, suggesting that the gene has conformed to the host genome.

Overall, these findings point to that MOQ_006101 was acquired by the lineage leading to Tcm rather than lost in T. c. cruzi and demonstrates an example of horizontal gene transfer between a photosynthesizing organism and a protozoan parasite. Interestingly, a genome comparison of two strains of the protozoan Giardia intestinalis also identified a strain-specific acetyltransferase [28]. Finally, the biological function of MOQ_006101, if any, remains to be determined.

Comparison of synteny reveals putative rearrangements

T. c. cruzi has previously been reported to exhibit extensive DNA content and karyotype variability [24,41-43]. We investigated sequence co-linearity of the assembled data and compared with the current chromosome-level assembly of Tcc CLBR. Scaffolds >25 kb were extracted from the assemblies, which resulted in 307 and 229 scaffolds for Tcm and Tcc X10 respectively, amounting to 50.7% (19.6 Mb/38.6 Mb) and 44.7% (19.4 Mb/43.4 Mb) of the genomes. In order to identify putative inter- and intra-chromosome rearrangements, scaffolds were queried against Tcc CLBR (non-Esm) using the alignment program promer [44]. The number of chromosome hits per scaffold was plotted and the results were inspected.

Table 4 List of hits obtained from PSI-BLAST after 4 iterations querying MOQ_006101 against GenBank non-redundant database

<table>
<thead>
<tr>
<th>Species</th>
<th>Description</th>
<th>Accession</th>
<th>CDD hit *</th>
<th>% Identity</th>
<th>BLAST E-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Populus trichocarpa</td>
<td>Predicted protein</td>
<td>XP_002298511.1</td>
<td>Cas1_AcylT</td>
<td>38%</td>
<td>0</td>
</tr>
<tr>
<td>Arabidopsis thaliana</td>
<td>Putative O-acetyltransferase</td>
<td>NP_586662.1</td>
<td>Cas1_AcylT</td>
<td>39%</td>
<td>0</td>
</tr>
<tr>
<td>Arabidopsis thaliana</td>
<td>AT5g46340/MPL12_14</td>
<td>AAL11600.1</td>
<td>Cas1_AcylT</td>
<td>38%</td>
<td>0</td>
</tr>
<tr>
<td>Arabidopsis thaliana</td>
<td>O-acetyltransferase-like protein</td>
<td>NP_180988.3</td>
<td>Cas1_AcylT</td>
<td>37%</td>
<td>0</td>
</tr>
<tr>
<td>Populus trichocarpa</td>
<td>Predicted protein</td>
<td>XP_002317300.1</td>
<td>Cas1_AcylT</td>
<td>37%</td>
<td>0</td>
</tr>
<tr>
<td>Vitis vinifera</td>
<td>CAS1 domain-containing protein 1-like</td>
<td>XP_002272126.2</td>
<td>Cas1_AcylT</td>
<td>38%</td>
<td>0</td>
</tr>
<tr>
<td>Arabidopsis lyrata subsp. Lyrata</td>
<td>O-acetyltransferase family protein</td>
<td>XP_002879497.1</td>
<td>Cas1_AcylT</td>
<td>37%</td>
<td>0</td>
</tr>
<tr>
<td>Arabidopsis lyrata subsp. Lyrata</td>
<td>Hypothetical protein</td>
<td>XP_002863407.1</td>
<td>Cas1_AcylT</td>
<td>39%</td>
<td>0</td>
</tr>
<tr>
<td>Ricinus communis</td>
<td>O-acetyltransferase, putative</td>
<td>XP_002519732.1</td>
<td>Cas1_AcylT</td>
<td>38%</td>
<td>0</td>
</tr>
<tr>
<td>Glycine max</td>
<td>CAS1 domain-containing protein 1-like</td>
<td>XP_003532649.1</td>
<td>Cas1_AcylT</td>
<td>38%</td>
<td>0</td>
</tr>
</tbody>
</table>

* The best hit from the NCBI Conserved Domain Database.
A total of 73 (Tcm) and 114 (Tcc X10) scaffolds contained hits to more than one chromosome from Tcc CLBR. However, manual examination showed that the vast majority of these hits were to gene family members (e.g. DGF, trans-sialidase, TcMUCII mucin) or other repeats. Hence, these were not likely to be rearrangements between chromosomes. 4 scaffolds were identified from Tcm (244, 732, 1101 and 2169) and 6 from Tcc X10 (94, 737, 1353, 2784, 2065 and 2359) that were involved in inter-chromosome rearrangements (Additional file 7: Figure S5). Moreover, Tcm and Tcc X10 both contained rearrangement in a region on chromosome 34, containing a repeat-like composition in Tcc CLBR. Scaffold 1101 from Tcm aligned with chromosomes 34 and 27. Scaffold 94 from Tcc X10 aligned with chromosomes 34 and 12. Also, scaffold 732 from Tcm aligned with the distal parts of Tcc CLBR chromosomes 22 and 42. In Tcm, VIPER elements were frequently found in regions where synteny was discontinued. Regions where rearrangements had occurred were frequently found inside unidirectional gene clusters.

Intra-chromosome rearrangements were searched for using the same strategy. This identified 23 and 13 scaffolds in Tcm and Tcc X10 respectively, where intra-chromosome rearrangements were identified. Frequently, one or several genes were found to have shifted location and were found to be located distally on the same chromosome. In a few cases, a certain structural variant was present in Tcm and Tcc X10 but not in Tcc CLBR, suggesting that it was introduced in the lineage leading to Tcc CLBR. Tcm scaffold 836 contained a large inverted region, flanked by VIPER elements. This inversion causes disruption of a head-to-head strand switch region. The larger number of structural rearrangements in Tcm likely reflects its phylogenetic distance from T. cruzi.

PCR validation was performed in order to validate the accuracy of the assembly and some of the identified rearrangements. Representative regions were selected from Tcm, Tcc X10 and Tcc CLBR and targeted for PCR amplification. The size of the PCR product was compared with the in silico expected size and confirm assembly consistency. In total, 3 of 4 genomic regions were...
successfully amplified from *Tcm*, 2 of 2 from *Tcc* X10 and 1 of 2 from *Tcc* CLBR (Additional file 8: Figure S6). Of which *Tcm* yielded the following PCR product sizes: 4, 0.5 and 0.8 kb. The first and second PCR products spanned across assembly gaps and therefore did not allow estimation of the expected sizes, but confirmed contigs to be accurately linked together. The third PCR reaction from *Tcm* resulted in the expected product size of 0.8 kb. *Tcc* X10 resulted in PCR products of sizes 0.8 kb and 1 kb, which were expected. The *Tcc* CLBR reaction resulted in the expected product size of 3 kb. Two PCR reactions did not work, which could either be due to non-optimal PCR conditions, formation of primer-secondary structures/duplexes or misassembly.

It is important to note that the present analyses are limited by the sequence continuity of the particular scaffold and therefore the presented numbers of observed rearrangements are likely to be underestimates. In conclusion, the majority of analyzed genomic regions from *Tcm* and *Tcc* X10 exhibited conserved local synteny with *Tcc* CLBR. However, insertions, deletions or other types of structural alterations occasionally interrupted synteny. These observations suggest that different *T. c. cruzi* lineages contain distinct karyotypes and other types of structural features, which have been fixed in a certain lineage. The cause of these rearrangements could either be due to random processes, i.e. oxidative stress or mistakes introduced by spontaneous cellular processes or perhaps less likely, physiological processes. Clearly, the presence of genetic variation other than SNPs provides an additional layer of complexity to studies of *T. c. cruzi* genetic variability.

### Widespread occurrence of copy number variation in *Tcm* and *Tcc* X10

Copy number variation has been reported from *T. c. cruzi* strains [33,34,45]. Such variation may represent important strain-specific characteristics, yet little is known about how *T. cruzi* lineages differ in this aspect. In the present study we investigated the occurrence of copy number variation in *Tcm* and *Tcc* X10 using short read depth (coverage). RT-qPCR was used to confirm some of the identified variations. Illumina reads were aligned to scaffolds >50 kb and a sliding window analysis was used to identify regions which exhibited higher than average coverage. Coverage was calculated in 100 bp windows with 50 bp overlap, i.e. the coverage of each position in the window was summed and log_{10}-scaled. The baseline coverage was determined for each genome and was used to assess if a region displayed elevated coverage compared to the rest of the genome. The baseline was calculated as the median and median absolute deviation (mad) of log_{10}-scaled coverage from all windows of one genome. This resulted in the baseline coverage (median ± mad) of 3.39 ± 0.35 and 3.39 ± 0.33 for *Tcm* and *Tcc* X10 respectively. A duplicated region was defined as a stretch of 1,000 bp containing at least 5 windows above 2X the median standard deviation from the baseline. If two or more stretches were adjacent to each other, these were merged and counted as one region. This identified 142 and 182 duplicated regions in *Tcm* and *Tcc* X10. The duplicated region was not always restricted to one gene. On scaffold 1093 (*Tcm*), the amplified region was 6 kb and contained four coding sequences, including a nucleoside transporter and a dynein gene (Figure 6). The same region was also found amplified in *Tcc* X10 (scaffold 1531; Figure 6). Interestingly, a nucleoside transporter has been implicated in drug resistance in *Trypanosoma brucei* [46]. Housekeeping-genes were also found amplified, for example, parafflagellar rod protein 3 (MOQ_003131) from *Tcm* gave rise to a CNV signal. RT-qPCR with primers targeting this gene resulted in Ct=13.60, compared with Ct=15.3 for 8-oxoguanine DNA glycosylase (MOQ_000430), which lacked a CNV signal. Moreover, the prostaglandin F2 alpha synthase gene (MOQ_004364) gave rise to a CNV signal, and RT-qPCR resulted in Ct=12.41. In *Tcc* X10, one example of CNV is in the epsilon tubulin gene (TCSYLVIO_007352), for which RT-qPCR resulted in Ct=9.29. Surface antigens were frequently found amplified; a mucin-like gene on scaffold 1070 (*Tcm*), a surface protease GP63 on scaffold 1108 (*Tcm*). Scaffold 1109 (*Tcm*) contains an amplification of cystathionine beta-synthase, scaffold 1420 (*Tcm*) contains an amplification of NAD(P)-dependent steroid dehydrogenase, scaffold 143 (*Tcm*) contains an amplification of ferric reductase. There was also evidence of differential copy number variation, suggesting amplification in one genome but not the other. One example is the amplification of a pyruvate phosphate dikinase on scaffold 1101 in *Tcm*, which does not give rise to a CNV signal in *Tcc* X10.

In order to identify chromosomal aneuploidies, we calculated the baseline coverage for each scaffold. Scaffolds with a median <3.2 and median absolute deviation <0.7 were extracted (empirically determined thresholds). This identified 7 scaffolds in *Tcm* with a lower average coverage: 950, 938, 79, 78, 70, 2392 and 2744. These genomic regions were homologous with chromosomes 25, 25, 12, 12, 12 and 7 in *Tcc* CLBR. No scaffolds with low overall coverage were identified in *Tcc* X10. This suggested the existence of monosomic chromosomes in *Tcm* but not in *Tcc* X10. As expected, heterozygosity was absent in these putatively monosomic *Tcm* regions. However, the homologous region in *Tcc* X10 displayed heterozygosity. This further supported the likely monosomic state of these regions in *Tcm*. Genomic qPCR with a primer pair targeting scaffold 78 in *Tcm* resulted in Ct=17.81 for the putative monosomic scaffold, whereas
for putative disomic regions on scaffold 1093 and 1101. Ct was 15.08 and 15.30. Moreover, we searched for evidence of higher ploidy levels. Scaffolds with median > 3.5 were extracted. This identified 14 and 5 scaffolds in Tcm and Tcc X10 respectively, with an increased overall coverage. These scaffolds showed homology with large mega-base chromosomes from Tcc CLBR, suggesting that higher ploidy levels may be more common in larger chromosomes.

The presented analysis confirms that copy number variation is a common feature of the Tcm and Tcc X10 genomes. In theory, copy number variation would not be beneficial for the parasite as it increases the amount of DNA that needs to be replicated and the energy cost of the cell. The evolutionary benefit of having such an excessive amount of genes would seem to be limited. It is possible that copy variation does not infer any evolutionary advantage for the parasite, but is only a consequence of sloppy or non-perfect DNA replication mechanisms of these parasites.

**Retrotransposons, repetitive elements and low complexity repeats**

Transposons are present in most eukaryotes and contribute to genome size and plasticity [47]. Trypanosomatid
genomes contain several families of dead and presumably active retrotransposons [48]. *Tcm* and *Tcc* X10 were searched for 11 classes of repetitive elements, including retrotransposons. 6.5% (2,344,982 Mb/34,233,090 Mb) and 9.9% (3,852,782 Mb/38,598,156 Mb) of the assembled bases corresponded to repetitive elements in *Tcm* and *Tcc* X10 respectively (Table 5). 8 of 11 repeat categories were more abundant in *Tcc* X10, with a total expansion factor of 1.26 in this genome compared with *Tcm* (8.2%/6.5%, Table 5).

The Long Terminal Repeat (LTR)-like retroelement VIPER [49] belongs to the superfamily tyrosine recombinase retrotransposons [50] and was the most abundant element in *Tcm* and *Tcc* X10 respectively, representing 24.5% (574,697 Mb/2,344,982 Mb) and 28.9% (1,116,378 Mb/3,852,782 Mb) of the repetitive elements (Table 5). The large amount of sequence related to these elements suggested that large-scale proliferation occurred before the split of *T. c. cruzi* and *T. c. marinkellei*. Furthermore, phylogenetic reconstruction based on a multiple sequence alignment (MSA) indicated some substructure between *Tcc* X10 and *Tcc* CLBR, whereas *Tcm* in large formed a more distant clade (Additional file 9: Figure S7). 69% (3,450/4,968) of the MSA sites were too diverged to be included in the multiple-alignment, suggesting that these elements have been inactive for a substantial time. The human infecting lineage does contain a larger amount of these elements, possibly due to a loss of VIPER-related sequences in *Tcm*. 3 repeat categories were on the contrary expanded in *Tcm*; the 2 low-abundance repeats NARTc and C6 and the abundant DIRE (degenerate Ingi/L1Tc-related retroelement) element.

*De novo* repeat discovery was performed in order to identify unique sequence repeats, using the program RepeatScout [51] and RepeatMasker [52]. RepeatScout identified 2,225 (*Tcm*) and 2,263 (*Tcc* X10) repeats of variable lengths. These repeats were then filtered using these criteria; i) removal of repeats shorter than 50 bp, ii) removal of repeats containing more than 50% low complexity sequence, iii) removal of repeats with fewer than 10 genomic copies, iv) removal of known repeats (i.e. present in *Tcc* CLBR). This decreased the number of hits to 20 (*Tcm*) and 3 (*Tcc* X10) using the outlined criteria. Manual examination of the *Tcm* repeats revealed that 12 corresponded to diverged *Tcc* CLBR sequences, including a spliced leader sequence and sequences related to MASP and TcMUCII mucin. We searched for these 8 repeats in the genome assemblies of *Tcc* X10 and *Tcc* CLBR as well as in raw reads, which decreased the number of *Tcm*-specific repeats to 7. The length of the identified *Tcm*-specific repeats varied between 60 to 896 bp, and BLAST searches resulted in non-significant hits to sequences of metazoan origin. These repeats were found exclusively on short contigs (0.5-1 kb), corroborating the idea that the repetitive components of these genomes have evolved faster. We estimated the copy number of the two longest repeats, Tcm-Rep1 (825 bp) and Tcm-Rep2 (896 bp) from the depth of 454 read coverage mapped on these sequences. The average 454 read coverage (12x) was then used to estimate copy number. The average read coverage was 1974 reads/position for Tcm-Rep1 and 1,494 reads/position for Tcm-Rep2. Hence, the

### Table 5 Comparison of repetitive elements

<table>
<thead>
<tr>
<th>Element</th>
<th><em>T. c. marinkellei</em></th>
<th><em>T. c. cruzi</em> Sylvio X10</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td># bp *</td>
<td>% Short reads b</td>
</tr>
<tr>
<td>VIPER</td>
<td>574,697 (1.679 %)</td>
<td>1.53</td>
</tr>
<tr>
<td>DIRE</td>
<td>433,619 (1.267 %)</td>
<td>1.156</td>
</tr>
<tr>
<td>L1Tc</td>
<td>432,474 (1.263 %)</td>
<td>1.168</td>
</tr>
<tr>
<td>TcTREZO</td>
<td>382,416 (1.117 %)</td>
<td>1.024</td>
</tr>
<tr>
<td>E22</td>
<td>223,679 (0.653 %)</td>
<td>0.630</td>
</tr>
<tr>
<td>SIRE</td>
<td>176,724 (0.516 %)</td>
<td>0.497</td>
</tr>
<tr>
<td>SZ23</td>
<td>94,765 (0.277 %)</td>
<td>0.224</td>
</tr>
<tr>
<td>CZAR</td>
<td>18,338 (0.054 %)</td>
<td>0.104</td>
</tr>
<tr>
<td>NARTc</td>
<td>4,705 (0.014 %)</td>
<td>0.010</td>
</tr>
<tr>
<td>C6</td>
<td>2,944 (0.009 %)</td>
<td>0.006</td>
</tr>
<tr>
<td>TCSAT1</td>
<td>621 (0.002 %)</td>
<td>0.149</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>2,344,982 (6.851%)</strong></td>
<td><strong>6.503%</strong></td>
</tr>
</tbody>
</table>

* a The sum of masked base pairs in the assembly. The number inside parenthesis refers to the percentage of assembled bases.

* b The percentage of short reads that was mapped on these features.

* c SE=Significantly Enriched. Refers to if one genome contained significantly more of this gene family. The significance was determined from an empirical distribution of read depth differences from homologous regions of *Tcm* and *Tcc* X10, corrected for genome size. The empirical distribution was used to calculate a p-value.
estimated copy number became 164 and 124 for Tcm-Rep1 and Tcm-Rep2 respectively. Taken together, these two repetitive elements contribute ~250 kb of sequences to the Tcm genome and also represent a large set of putative Tcm-specific sequences. Since the repeats were not found in Tcc X10 and Tcc CLBR, it is possible that a loss has occurred in the lineage leading to the human infective T. c. cruzi.

T. c. marinkellei invades non-bat epithelial cells in small numbers and divides intracellularly

Experimental infections were performed on three mammalian cell lines to further understand the potential of Tcm to invade non-bat derived cells. The following lines of epithelial cells were used; Vero cells (kidney cells from African green monkey), OK cells (from a North American opossum) and Tb1-lu cells (bat lung). Tcm metacyclic trypomastigotes were incubated overnight with cells from each cell line (Materials and Methods). Extra- and intracellular parasites were immunolabelled using Tcm and Tcc positive sera and anti-whole cell body antibody (Figure 7). In parallel, intracellular parasites were stained with Giemsa dye. Both experiments independently showed that Tcm is capable of invading each of the three cell lines. Tcm did not show a particular preference for the bat cell line.

We also investigated the ability of Tcm to replicate intracellularly using the same cell lines. The human infective Tcc was included as a positive control. Infected cells were incubated over a 5 day time course, the development of intracellular amastigotes during this period indicated that Tcm is capable of intracellular proliferation (Figure 7). Amastigogenesis and amastigote proliferation of Tcm following metacyclic invasion appeared to be analogous to Tcc controls. These data suggested that Tcm is capable of infecting other cells than strictly those from bats, and hence, that the infection is not blocked by species-specific host cell tropism mediated at the level parasite entry. In conclusion, the infection barrier must therefore arise in a different context, though whether this relates to different aspects of Tcm specific biology or as physiological or immunological differences between hosts, or as a combination of both, remains to be elucidated.

Conclusions

This study is the first genome analysis of a non-human associated member of Schizotrypanum. Our aim was to identify genome sequence differences that may relate to host specificity or other phenotypical differences, as well as to further understand the evolution of these parasite lineages. We found a slightly smaller genome of T. c. marinkellei compared with the human infective strains, although it remains an open question if this is a general tendency among bat-associated trypanosomes. T. c. marinkellei and T. c. cruzi shared the same set of core genes, i.e. there were no missing coding sequences in terms of housekeeping genes. On the contrary, several gene families were expanded in T. c. cruzi Sylvio X10, contributing to the larger genome size. This suggested that T. c. cruzi Sylvio X10 have a more versatile toolbox of surface antigens, which may reflect an adaptation to its host. Interestingly, one subspecies specific acetyltransferase gene was identified in T. c. marinkellei, containing detectable homology with genes from photosynthesizing organisms. It appears likely that this gene was acquired after the split of T. c. cruzi and T. c. marinkellei, since the gene was missing from T. c. cruzi strains Sylvio X10 and CL Brener. The gene represents a rare example of gene transfer between distantly related eukaryotes and may provide additional functionality to T. c. marinkellei. Future efforts will be required to understand its function. Considering the divergence time between T. c. marinkellei and T. c. cruzi (~6.5-8.5 MYA [10-12]), remarkably few absolute gene differences were present. This suggests that the core gene content of T. cruzi lineages is relatively stable, whereas the repetitive component is allowed to undergo more rapid changes. The low number of subspecies specific genes suggests that phenotypic variation, like host specificity, might be encoded by more discrete variation, e.g. via non-synonymous nucleotide variants leading to specific protein isoforms. The difficulty to explain how the genome encodes phenotypes like host-specificity is further illustrated by our finding that T. c. marinkellei invades non-bat cells, which indicates that the machinery to invade host cells is functionally conserved. The two subspecies T. c. marinkellei and T. c. cruzi were on average ~7.5% diverged in coding sequences with respect to single nucleotide differences. The large number of small nucleotide differences may have implications on phenotypic variation via the formation of new alleles. The present study has provided many new candidate genes, including putative antigens that can provide starting points for functional investigation of phenotypic variation of these parasite lineages.

Extensive copy number variation of various genes was identified. Copy number variation has been suggested as means for the parasite to increase gene expression in the absence of transcriptional regulation. These findings are not surprising and are corroborated by the longstanding knowledge of genomic variability in T. c. cruzi [24,33,41-43]. It is possible that phenotypes may be encoded at the transcriptional level. Interestingly, the T. c. marinkellei and T. c. cruzi genomes also contained variation in the amounts of non-coding repeats, related to retroelements and other previously uncharacterized repeats. As these differences were substantial, it remains plausible that whole chromosomes or chromosomal
chunks have been lost in *T. c. marinkellei*. None of the larger chromosomes were missing, suggesting that smaller chromosomes harboring surface antigens or other repeats have been lost. The plasticity of the *T. cruzi* karyotype is further demonstrated by the fact that certain chromosomes appear to be monosomic in one subspecies but not in the other. The existence of such monosomic chromosomes reduces allelic redundancy and might have implications on transcript abundance. Karyotype variability therefore stands as another possible source of phenotypic variation. Finally, the amount of intraspecific genetic variation identified in this study is likely to

**Figure 7 Cell invasion assay.** (A) Intracellular *T. c. marinkellei* parasites stained with Giemsa. Scale bars correspond to: 10 μm (field) and 5 μm (enlarged). (B) Immunolabelled intracellular and extracellular *T. c. marinkellei* parasites. Intracellular parasites were labelled with anti-WCB antibody (green), while extracellular parasites were labelled with anti-WCB antibody (green) and anti-*T. c. marinkellei* serum (red), which superimposed gives the yellow color. Nuclei and kinetoplasts were counter stained in DAPI (blue). (C) Number of intracellular *T. c. marinkellei* parasites in the Giemsa assay. (D) Number of intracellular *T. c. marinkellei* parasites in the immunolabeling assay. (E) Intracellular *T. c. cruzi* and *T. c. marinkellei* parasites in three different cell types. *T. c. cruzi* and *T. c. marinkellei* parasites were incubated for 5 days with Vero (monkey), OK (opossum) and Tb1-Lu cells before Giemsa staining. Two hundred cells were assayed in 3 independent experiments for Giemsa and immunolabeling assays. The scale bars correspond to 5 μm.
represent only the tip of the iceberg in terms of the actual genetic variation present in natural reservoirs.

Taken together, the presented draft genomes raise further questions about genome evolution and diversity in this group of protozoa, and the putative functional implications of this variation. Further exploration of the genetic diversity within *Schizotrypanum* should therefore be a future priority as this may help to resolve complex relationships between parasites, vectors and hosts. The decreasing cost and time for whole genome sequencing should therefore pave the way for further large-scale efforts to understand the underlying genetic basis of these parasites.

**Methods**

**Accession numbers**

Sequence data and annotations have been deposited in NCBI GenBank under the accession numbers AHKC01000000 (*T. c. marinkellei* B7) and ADWP02000000 (*T. c. cruzi* Sylvio X10). The data can also be downloaded from http://www.ki.se/chagasepinet/genomes.html.

**Cell culture, library preparation and sequencing**

*T. c. marinkellei* B7 clone 11 and *T. c. cruzi* Sylvio X10 clone 1 were cultured using standard conditions (supplemented RPMI + 10% FBS). Parasite DNA was extracted using the Puregene kit. Genome size estimation of *T. c. marinkellei* was performed using flow cytometry as described by Lewis et al. [24]. Illumina sequencing: The *Tcm* and *Tcc* X10 mate pair libraries were prepared according to Van Nieuwenburgh et al. [22]. Initially, the paired-end protocol for 3 kb libraries from Roche/454 was used until circularized. After circularization, the libraries were prepared following the Illumina mate-pair protocol: 5 µg of genomic DNA was sheared to approximately 3 kb and end-polished. Fragments were then ligated to the Roche/454 circularization adapters and size selected using AMPure beads (Beckman Coulter). The ligated DNA was circularized using Cre-recombinase and then exonuclease treated. All enzymes were obtained from New England Biolabs. The circular DNA was fragmented using Covaris S2, end-repaired and purified using streptavidin coated magnetic beads. The DNA was then A-tailed and ligated with sequencing adapters and PCR amplified. The post-circularization steps were performed using the reagents either supplied or referenced by the Illumina mate-pair protocol. The clustering was performed on a cBot cluster generation system using a HiSeq paired-end read cluster generation kit. The samples were sequenced on an Illumina HiSeq2000 as 2x100-nt reads (one lane for each genome). Base conversion was done using Illumina OLB v1.9. 454 sequencing: Prepared according to the manufacturer’s instructions and sequenced on a 454 instrument with Titanium chemistry.

**Sequence assembly**

Illumina reads were quality filtered and trimmed using the fastq_quality_filter program of the fastx toolkit (parameters: -q 20 -p 95). Illumina reads were assembled with Velvet v1.1.04 [53], using the empirically determined kmer length of 43 and a minimum contig length of 500 bp. *velvet* and *velvetg* were called with the following commands, respectively: `<name> 43 -fastq -shortPaired1 input.fastq' and `<name> -min_contig_lgh 500 -exp_cov auto -ins_length 2000 -ins_length_sd 2000 -amos_file no -scaffolding no -unused_reads yes'. Unused reads were extracted and subjected to a second round of Velvet assembly using a kmer length of 53 (empirically determined) and a minimum contig length of 400 bp (*velvetg*: `<name> -min_contig_lgh 400 -exp_cov auto -ins_length 10 -ins_length_sd 5000 -scaffolding no`). 454 reads were assembled with CELERA v6.1 [54] (default settings). 454-related insertion-deletion errors in the assembly were corrected using the Illumina reads: Illumina reads were aligned with bwa [55], and the resulting SAM file was then scanned in order to identify indels. In order to correct a position, at least 10 Illumina reads were required to support the change. The three assemblies (i. Illumina Velvet 1; ii. Illumina Velvet 2; iii. 454 CELERA) were pooled and merged into a non-redundant assembly. Assembly merging was performed using the Zorro pipeline [56], relying on minimus2 and mummer to split and merge contigs. The merged assembly was filtered to include only contigs longer than 500 bp. Maxicircle (mitochondrial) sequences were identified using megablast and removed. Minicircle (mitochondrial) sequences were identified and removed by searching for the TCKIN2 signature sequence. Distance information from mate-pairs were used to order and orient contigs into scaffolds. The software SSPACE was used for scaffolding [57]. A small number of intra-scaffold gaps (~200genome) could be closed using the overlap between adjacent contigs. The final assembly was subjected to gap closure using the IMAGE pipeline [23].

**Bioinformatics analyses**

Annotation: Gene models were transferred from *Tcc* CLBR using Perl scripts, and additional genes were called using GeneMarkS [58]. Annotations were manually curated using the Artemis Comparison Tool [38]. Orthologous genes were identified using the best reciprocal BLASTp hit (E-value 1e-10). Unique genes were identified using BLASTp and tBLASTn searches. Genes in synteny were determined using homology of surrounding genes. At least one adjacent homologous gene was required to be present in order to call a gene
Amplicons were visualized using a 1.3% agarose gel son with the other amplicons that uses just 10 seconds. For 3 kb size uses a 45 seconds extension step, in comparison step at 72°C for 2 minutes. The Tc_CLB1 amplicon for 10 seconds and 72°C for 10 seconds and a final extension for 2 minutes, 35 cycles of 98°C for 10 seconds, 60°C for 60 seconds. The cycling conditions were as follows: Initial denaturation at 98°C for 2 minutes, 35 cycles of 98°C for 10 seconds, 60–68°C for 10 seconds and 72°C for 10 seconds and a final extension step at 72°C for 2 minutes. The Tc_CLB1 amplicon of 3 kb size uses a 45 seconds extension step, in comparison with the other amplicons that uses just 10 seconds. Amplicons were visualized using a 1.3% agarose gel stained with ethidium bromide. Quantitative Real-time PCR (RT-qPCR) on the specific acetyltransferase gene in *Tom*: RNA extraction was performed using the RNeasy Mini kit (Qiagen). RNA was converted to cDNA with reverse transcriptase and random hexamer primers. Reactions were performed using Power SYBR Green MasterMix (Life Technologies) under standard conditions. Template concentration was 50 ng/μl and 1 μl of template was used in each reaction. Primer concentration was 0.2 μM in 20 μl of final volume. Each experiment was performed in triplicate and the average cycle threshold (Ct) value was used as a measurement of initial template abundance. All reactions were performed on an ABI 7300 Real-time PCR system. The following primer pairs were selected for the experiment (5’ to 3’; unique gene: TTGCAGCATATGTTGGATG (F), ACGT TTAAAGAAAACGGCTGCT (R), hypoxanthine-guanine phosphoribosyltransferase: GCCCTATGTCACCCCTT (F), AAGACGTGACCTAACCCA (R), 18S rRNA: TTACGTCTGCCATTTGTA (F), TTCGGTCAAGT GAAGCACTC (R). RT-qPCR to validate copy number variation: Experimental conditions were similar as for the previous experiment, except that genomic DNA was used (20 ng/μl, 1 μl loaded).

Cell invasion assay

Vero cells were maintained in DMEM + glutaMAX (Gibco, Invitrogen, UK) supplemented with 10% fetal calf serum (PAA laboratories, UK), 5mM l-glutamine, 50 μg/ml streptomycin and 50 units/ml penicillin. Tb1 lu cells (HPACC, UK) were maintained in MEM (Sigma-Aldrich, UK) and supplemented as described above. OK cells (HPACC, UK) were maintained in MEM (Sigma-Aldrich, UK) with 10% fetal calf serum, 5 mM l-glutamine, 50 μg/ml streptomycin and 50 units/ml penicillin and 5% non-essential amino acids. *T. c. marinkellei* epimastigotes from lineage B7 cl11 were grown in Liver Infusion Tryptose (LIT) and *T. c. cruzi* strain M6241 was grown in RPMI for 2 weeks prior to experiments. Each cell line was seeded at a density of 10⁵ per ml onto 13 mm diameter coverslips and allowed to grow overnight. Cells were then washed and the growth media replaced with media containing 10⁵ metacyclic *T. c. marinkellei* and incubated at 37°C for either overnight or 5 days. The media was removed and cells were washed twice with PBS and either fixed with 4% paraformaldehyde for 20 min or ice cold methanol. Methanol fixed cells were stained with Giemsa for 10 min and imaged with a Zeiss Axioskop 2 microscope and a Zeiss AxioCam Hrc camera. Paraformaldehyde fixed cells were blocked in 10% goat serum and labelled with *T. c. marinkellei* positive serum for 1 h before incubating with AffiniPure Fab fragments (Stratagene Scientific Ltd., UK) for 10 min, these epitopes were then recognised by anti-rabbit Alexa Fluor 568. The cells were blocked again in 10% goat serum and permeabilised.
in 1% NP40 for 3 min before labelling with anti-whole cell body (WCB) antibody [72] (kindly provided by Prof. K. Gull) for 1 h recognised by anti-mouse Alexa Fluor 488 and finally DAPI stained before mounting in Fluoromount (Sigma-Aldrich, UK). Antibody labelled cells were visualised by a Zeiss Axiosplan 2 microscope and Zeiss AxioCam MRm camera all image processing was done with Axiovision 4.7 software. Two hundred cells were assayed in the overnight experiments and the results are expressed as an average of three independent experiments.

Additional file

Additional file 1: Figure S1. Flow cytometry analysis of the T. c. marinkellei genome size. Description: Fluorescence emission histograms for propidium iodide-labelled epimastigotes showing relative DNA contents of T. c. cruzi Esm3/ Tcll, T. c. cruzi Sylvio X10/ Tcl and T. c. marinkellei B7/11.

Additional file 2: Figure S2. Histogram and smoothed density estimate of assembly-wide coverage differences between Tcm and Tcc X10. Description: (A) Histogram of percentage short read coverage differences from homologous regions. Percentages have been corrected for genome size. Vertical red lines indicate the lower and upper 2.5% quantiles. (B) Smoothed kernel density estimate of the left histogram created using logspine R package.

Additional file 3: Figure S3. Sequence variation of the TcMUCII mucin gene family. Description: Entropy plots of the TcMUCII mucin gene family. TcMUCII mucin genes were extracted from Tcm, Tcc X10 and Tcc CLBR non-Esm. Sequences were aligned with ClustalW v2.1. Sequence entropy was calculated using the entropy function of the R package bio3d. Only alignment positions with less than 10% gaps were included in the analysis. The normalized entropy score was then plotted as a function of alignment position, where conserved sites (low entropy) score 1 and diverse (high entropy) sites score 0. The analysis indicated that 5' and 3' termini of TcMUCII mucin genes generally are the most conserved in all three genomes and that the central region is the most variable.

Additional file 4: Table S1. Maxicircle gene coordinates and metrics. Description: Gene metrics for T. cruzi and T. c. marinkellei maxicircles. Including coordinates, average identity and length.

Additional file 5: Figure S4. Maxicircle phylogenetic tree. Description: Maximum likelihood phylogenetic tree of the maxicircle sequences from T. c. marinkellei, T. c. cruzi Sylvio X10, T. c. cruzi CL Brener, T. c. cruzi Esmeraldo using T. brucei and T. tarentolae as outgroups. The full maxicircle sequences were aligned with ClustalW v2.1 and the subsequent alignment was filtered using Gblocks (default settings). The tree was inferred using MEGA v5.1 from 13,731 (49%) alignment positions.

Additional file 6: Table S2. Ratio of non-synonymous and synonymous nucleotide substitutions. Description: Orthogonal gene pairs were used between T. c. marinkellei and T. c. cruzi CL Brener displaying elevated dN/dS (> 1.1). The yri00 program was used to calculate dN and dS.

Additional file 7: Figure S5. Disruption of sequence co-linearity. Description: Disruption of chromosomal co-linearity between T. c. marinkellei and T. c. cruzi CL Brener non-Esmeraldo-like (A) as well as between T. c. cruzi Sylvio X10 and T. c. cruzi CL Brener non-Esmeraldo-like (B). Black chromosomes prefixed with ‘Chr’ represent sequences from Tc CL Brener whereas white chromosomes prefixed ‘contig’ represent sequences from Tcm and Tcc X10 assemblies. Alignments were generated using the pomer software (Kurtz et al. 2004). Chromosomal stretches marked with green color represent gaps in the assembly. Only gaps larger than 5 kb are shown. The most outer numbers are sequence identifiers.

Competing interests
The authors declare that they have no competing interests.

Authors’ contributions
OF carried out the bioinformatics analyses and drafted the manuscript. CTL, SO, CEB, LAM, MDL, MSL carried out cell culture, flow cytometry, PCR experiments, cell invasion assays and participated in the bioinformatics analyses. BA, MAM, KMT, CJM conceived the study, and participated in its design and coordination and helped to draft the manuscript. All authors read and approved the final manuscript.

Acknowledgements
We would like to acknowledge Rickard Sandberg and UPMAX for providing computational resources and the Genomes Platform at SciLifeLab Stockholm for library preparation and sequencing. We also would like to thank the anonymous reviewers for their suggestions on how to improve the manuscript. This work was supported by the European Union Seventh Framework Program (http://cordis.europa.eu fp7 grant 223034 (ChagasEpiNet). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Author details
1Department of Cell and Molecular Biology, Karolinska Institutet, Box 285, Stockholm, SE 171 77, Sweden. 2Norwich Medical School, University of East Anglia, Norwich, Norfolk NR4 7TJ, United Kingdom. 3Department of Pathogen Molecular Biology, Faculty of Infectious and Tropical Diseases, London School of Hygiene and Tropical Medicine, Keppel Street, London, United Kingdom. 4Centro de Investigaciones en Microbiologia y Parasitologia Tropical, Universidad de los Andes, Santafe de Bogotá, Colombia.

Received: 2 February 2012 Accepted: 1 October 2012 Published: 5 October 2012

References


doi:10.1186/1471-2164-13-531
Cite this article as: Franzén et al.: Comparative genomic analysis of human infective Trypanosoma cruzi lineages with the bat-restricted subspecies T. cruzi marinkellei. BMC Genomics 2012 13:531.
COVER SHEET FOR EACH ‘RESEARCH PAPER’ INCLUDED IN A RESEARCH THESIS

Please be aware that one cover sheet must be completed for each ‘Research Paper’ included in a thesis.

1. For a ‘research paper’ already published

   1.1. Where was the work published? .................................................................

   1.2. When was the work published? .................................................................

   1.2.1. If the work was published prior to registration for your research degree, give a brief rationale for its inclusion
   N/A ..........................................................................................................................
   ..........................................................................................................................

   1.3. Was the work subject to academic peer review? .............................................

   1.4. Have you retained the copyright for the work? Yes/No

   If yes, please attach evidence of retention.

   If no, or if the work is being included in its published format, please attach evidence of permission from copyright holder (publisher or other author) to include work

2. For a ‘research paper’ prepared for publication but not yet published

   2.1. Where is the work intended to be published? ..............................................

   2.2. Please list the paper’s authors in the intended authorship order

   ..........................................................................................................................

   2.3. Stage of publication – Not yet submitted / Submitted / Undergoing revision from peer reviewers’ comments / In press

3. For multi-authored work, give full details of your role in the research included in the paper and in the preparation of the paper. (Attach a further sheet if necessary)

   The candidate participated in data assembly, analysis, interpretation and manuscript preparation.

   ..........................................................................................................................

NAME IN FULL (Block Capitals) LOUISA ALEXANDRA MESSENGER

STUDENT ID NO: 223021

CANDIDATE’S SIGNATURE ................................................................. Date 10/2/15

SUPERVISOR/SENIOR AUTHOR’S SIGNATURE (3 above) .................................................................
Molecular Epidemiologic Source Tracking of Orally Transmitted Chagas Disease, Venezuela

Maikell Segovia, Hernán J. Carrasco, Clara E. Martínez, Louisa A. Messenger, Anaíbeth Nessi, Juan C. Londoño, Raul Espinosa, Cinda Martínez, Mijares Alfredo, Rafael Bonfante-Cabarcas, Michael D. Lewis, Belkisyló A. de Noya, Michael A. Miles, and Martin S. Llewellyn

Oral outbreaks of Chagas disease are increasingly reported in Latin America. The transitory presence of Trypanosoma cruzi parasites within contaminated foods, and the rapid consumption of those foods, precludes precise identification of outbreak origin. We report source attribution for 2 peri-urban oral outbreaks of Chagas disease in Venezuela via high resolution microsatellite typing.

Rapid urbanization presents new challenges for Chagas disease control in Latin America. Foci of disease are now reported in slums surrounding several Andean cities (1–3). Oral transmission is believed responsible for recent outbreaks of Chagas disease, most of which were characterized by atypically severe symptoms (4,5). Many cases have occurred in urban settings (5,6), amplifying the size and effect of the outbreaks.

Sources of orally transmitted disease outbreaks vary, but contaminated food and juices are often blamed. However, after a contaminated food is eaten, it may take weeks for the onset of clinical signs and symptoms, and direct molecular and cytological incrimination of a particular batch of food/beverage has not been possible (5). Thus, evidence pointing to particular foodstuffs is often circumstantial.

Molecular epidemiologic analyses of human and environmental isolates are routinely used to track the source of outbreaks caused by foodborne pathogens. High-resolution molecular markers have been developed and validated for Trypanosoma cruzi, the parasite that causes Chagas disease (7,8). These markers, used in conjunction with careful sampling, can identify the source of foodborne outbreaks.

The Study

We studied 2 outbreaks of orally transmitted Chagas disease (120 cases, 5 deaths). The first occurred in Chichiriviche, Vargas State, a coastal community (population ≈800 persons) ≈50 km northwest of Caracas, Venezuela. The outbreak occurred at a primary school where food was prepared on site. In early April 2009, a total of 71 children (6–13 years of age) who attended the morning school shift and 14 adults became ill. Exposure of these persons to T. cruzi was established by use of IgM and IgG ELISA. Parasitemia was observed in 33 of the patients with serologic results positive for T. cruzi infection (9,10).

The second outbreak occurred in Antimano, a peri-urban slum southwest of central Caracas (online Technical Appendix Figure 1, wwwnc.cdc.gov/EID/article/19/7/12-1576-Techapp1.pdf). In May 2010, 35 patients with suspected T. cruzi infection were examined at Hospital Miguel Perez Carreno in Caracas. Patients reported that they regularly ate at the same local communal canteen. Among the patients tested, 15 were positive for T. cruzi IgM and IgG (9). Parasitemia in 14 patients was confirmed indirectly by hemoculture. Of the 35 patients, 21 (2 adults, 19 children) were hospitalized.

To enable outbreak source attribution, we undertook intensive additional sampling of contemporary, nonhuman sources local to each outbreak and of human and nonhuman sources from more distant localities throughout Venezuela. In total, 246 T. cruzi strains and clones were typed for 23 microsatellite markers (online Technical Appendix 1 Table) (8). A list of the samples and their sites of origin is in online Technical Appendix 2 (wwwnc.cdc.gov/EID/article/19/7/12-1576-Techapp1.xlsx).

Individual level sample clustering was defined first by constructing a neighbor joining tree based on pairwise distances between multilocus genotypes (Figure 1). A second analysis used K-means clustering and discriminant analysis of principal components (Figure 2) (11). To assess connectivity between human and nonhuman outbreak cases, pairwise genetic differentiation ($F_{ST}$) was calculated (Table 1). Population-level genetic diversity was assessed first by calculating allelic richness then private allele frequency...
over loci between each human–nonhuman population pair (Table 2). Geographic sampling distribution is shown in online Technical Appendix 1 Figure 2.

Clustering results determined by discriminant analysis of principal components and neighbor joining were broadly congruent. In the former, 19 clusters were defined; sample allocations are included in online Technical Appendix 2. Substantial overlap existed between some clusters, especially those from Capital and Vargas States (e.g., those labeled “x” in Figure 2), while others were highly distinct (e.g., cluster 8 in Figure 2). Human isolates from both oral outbreaks are extremely distinct from non-orally transmitted isolates collected from humans throughout Venezuela. Almost all these presumably vector-transmitted strains are closely related to one another, despite their geographic dispersal (cluster 8 in Figure 2). By comparison, oral outbreak strains that were isolated a mere 50 km apart (clusters 2, 5, 7, 9, 15–17) are far more globally diverse. Unlike most human isolates in Venezuela, which are distinct from nonhuman strains, oral outbreak isolates are interspersed among nonhuman strains from Venezuela. Furthermore, samples from both outbreaks clustered among nonhuman strains local to that outbreak, clearly indicating a local origin. Oral samples from each outbreak are polyphyletic with respect to strains from their immediate environment, a finding consistent with multiple contamination events or multiclonal infection sources.

\( F_{ST} \) values further support connectivity between outbreak and local environmental samples in both Antimano and Chichiriviche (Table 1). A lack of private alleles between human and nonhuman isolates also supports a local source for the Chichiriviche outbreak (Table 2). \( F_{ST} \) values in the 4-way comparison between outbreak and local environmental strains are, however, somewhat equivocal with respect to the entire dataset (Table 1). Cluster analysis showed that the human and nonhuman strains from Chichiriviche interspersed with strains from other states in Venezuela (Figure 2). Thus, we cannot confirm a uniquely local origin for the Chichiriviche outbreak, despite a low value for \( F_{ST} \), and it is possible that some contaminating strains originated elsewhere.

**Conclusions**

This study demonstrates the value of rigorous molecular epidemiologic analysis of orally transmitted *T. cruzi* outbreaks, including the importance of appropriate sampling to identify the origin of the infecting strains. The foodstuff that propagated the peri-urban outbreak in Antimano was certainly contaminated locally. An active nonhu-

<table>
<thead>
<tr>
<th>Isolate, location</th>
<th>Human isolates from</th>
<th>Nonhuman isolates from</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Antimano</td>
<td>Chichiriviche</td>
</tr>
<tr>
<td>Human Antimano</td>
<td></td>
<td>0.000</td>
</tr>
<tr>
<td>Chichiriviche</td>
<td>0.201</td>
<td></td>
</tr>
<tr>
<td>Nonhuman Antimano</td>
<td>0.093</td>
<td>0.170</td>
</tr>
<tr>
<td>Chichiriviche</td>
<td>0.088</td>
<td>0.053</td>
</tr>
</tbody>
</table>

*Lower left shows linearized \( F_{ST} \) (genetic differentiation) values; upper right shows associated p values.*
man transmission cycle in the slums of Caracas, maintained by Rattus rattus rodents and Panstrongylus geniculatus triatomines, is the likely source. The Chichiriviche outbreak, however, has potential sources both in and outside the immediate area. As found in Chagas disease outbreaks linked to açaí palm fruit in Brazil (12), the T. cruzi parasite can survive for several days in some foodstuffs (13). Also, triatomines can survive for months in harvested crops; thus, multiple hygiene interventions are potentially necessary along the food production line (14). Nonetheless, if the foodstuff implicated was prepared locally, local contamination represents the most likely source of the outbreak. Study of additional nonhuman strains from Chichiriviche is necessary to support this assertion.

Crucial to understanding parasite transmission in general, we believe, are genetic differences between strains from orally and non-orally transmitted human cases. All TcI strains appear to be infective to humans and adapted to long-term carriage (8). However, the presence of a common, reduced-diversity TcI genotype cluster (TcI\textsubscript{DOM}) among a high proportion of human Chagas disease cases in South America is also well established (7,8). We originally hypothesized that TcI\textsubscript{DOM} was maintained, despite the presence of sympatric and infective sylvatic strains, because of low parasite transmission efficiency by invasive sylvatic vectors (8). Oral transmission is likely to be much more efficient. Thus, unlike TcI\textsubscript{DOM} strains, those from orally transmitted T. cruzi cases demonstrated high genetic diversity and clearly originated from local nonhuman T. cruzi populations. However, it is also true that all TcI\textsubscript{DOM} strains we isolated originated from patients with chronic infection, and all orally transmitted cases were in the acute phase. We cannot, therefore, rule

### Table 2. Sample size corrected diversity between Trypanosoma cruzi isolates derived from humans and the local environment during an outbreak of orally transmitted Chagas disease in 2 areas of Venezuela*

<table>
<thead>
<tr>
<th>Isolate, location</th>
<th>No. isolates/no. genotypes</th>
<th>Sample size corrected allelic richness ± SE</th>
<th>Mean no. private alleles/locus ± SE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Human Antimano</td>
<td>30/26</td>
<td>2.735 ± 0.291</td>
<td>0.32 ± 0.113</td>
</tr>
<tr>
<td>Chichiriviche</td>
<td>12/9</td>
<td>3.459 ± 0.412</td>
<td>0</td>
</tr>
<tr>
<td>Nonhuman Antimano</td>
<td>107/91</td>
<td>2.946 ± 0.320</td>
<td>0.86 ± 0.203</td>
</tr>
<tr>
<td>Chichiriviche</td>
<td>13/13</td>
<td>3.443 ± 0.409</td>
<td>0</td>
</tr>
</tbody>
</table>

Figure 2. Discriminant analysis of principal components showing genetic clustering among Trypanosoma cruzi isolates from 2 outbreaks of oral disease in northern Venezuela. Six principal components were retained, explaining 80% of the diversity. Ellipses correspond to the optimal (as defined by the Bayesian information criterion minimum) number of population clusters among the genotypes analyzed. Images indicate sample host origin (human, rodent, marsupial, or triatomine), while colors correspond to the key in Figure 1. A full list of samples and population assignments (numbered boxes) is included in online Technical Appendix 2 (wwwnc.cdc.gov/EID/article/19/7/12-1576-Techapp1.xlsx). Dashed box indicates the isolates associated with the outbreaks.
out a role for immune selection in driving the frequency of TcI.DOM infections among humans; such selection represents an intriguing topic for future enquiry.

Molecular tools and reference datasets are now available to determine the source of acute Chagas disease outbreaks within days of their occurrence. The plummeting cost of such analyses means it is time to apply population genetic techniques and markers developed for trypanosomes as genuine epidemiologic tools.

Acknowledgments

Sincere thanks to Maria Dora Feliciangeli for her unconditional support. We also thank M. Rodriguez, L. Briceño, J. Ortegoza, M. Torres, F. Alfonso, H. Montañez, G. de Valenzuela, M. Sanchez, J. Lara, C. Viloria, J. Bracho, and L. Barrios for their support in the clinic, laboratory, and field.

This work was supported by FONACIT (grant G-2005000827) and the European Union (Fp7 grant 223034).

Mr. Segovia is a researcher at the Laboratorio de Biología Molecular de Protozoarios, Universidad Central de Venezuela, in Caracas, and he works in several clinical laboratories in the area. He has a keen practical interest in disease epidemiology and pathogen population genetics.

References


Address for correspondence: Hernan J. Carrasco, Universidad Central de Venezuela, Caracas, Venezuela; email: hernan.carrasco@ucv.ve
Please be aware that one cover sheet must be completed for each ‘Research Paper’ included in a thesis.

1. For a ‘research paper’ already published
   
   1.1. Where was the work published? .................................................................
   
   1.2. When was the work published? .................................................................
   
   1.2.1. If the work was published prior to registration for your research degree, give a brief rationale for its inclusion
   ..........................................................................................................................
   ..........................................................................................................................
   ..........................................................................................................................

   1.3. Was the work subject to academic peer review? ...........................................
   ..........................................................................................................................
   ..........................................................................................................................
   ..........................................................................................................................

   1.4. Have you retained the copyright for the work? Yes / No
   If yes, please attach evidence of retention.
   If no, or if the work is being included in its published format, please attach evidence of permission from copyright holder (publisher or other author) to include work

2. For a ‘research paper’ prepared for publication but not yet published

   2.1. Where is the work intended to be published? Emerging Infectious Diseases
   ..........................................................................................................................

   2.2. Please list the paper’s authors in the intended authorship order


   2.3. Stage of publication – Not yet submitted (Submitted) Undergoing revision from peer reviewers’ comments / In press

3. For multi-authored work, give full details of your role in the research included in the paper and in the preparation of the paper. (Attach a further sheet if necessary)
   The candidate was responsible for generating, assembling and analyzing all mitochondrial sequence data. The candidate also participated in overall data interpretation and drafting of the final manuscript.

NAME IN FULL (Block Capitals) LOUISA ALEXANDRA MESSENGER

STUDENT ID NO: 223021

CANDIDATE’S SIGNATURE................................................................. Date 10/2/15

SUPERVISOR/SENIOR AUTHOR’S SIGNATURE (3 above)..........................................................................................................

Improving health worldwide www.lshtm.ac.uk
Trypanosoma cruzi I Chronic Chagasic Cardiomyopathy and Chagas Disease Reactivation in Boston, Massachusetts, USA

Running Title: T. cruzi I Chagas heart disease/ reactivation in Boston

Jaime A. Costales¹, Camille N. Kotton, Andrea C. Zurita-Leal², Josselyn E. Garcia-Perez³, Martin S. Llewellyn⁴, Louisa A. Messenger, Tapan Bhattacharyya, Barbara A. Burleigh

Author affiliations:
Pontificia Universidad Católica del Ecuador, Quito, Ecuador (J.A. Costales, A.C. Zurita-Leal, J.E. García-Pérez), Massachusetts General Hospital, Boston, USA (C.N. Kotton), London School of Hygiene and Tropical Medicine, London, UK (M.S. Llewellyn, L.A. Messenger, T. Bhattacharyya), Harvard School of Public Health, Boston, USA (B.A. Burleigh).

Current affiliate: 2 University of Glasgow, Glasgow, Scotland, UK; 3 Katholieke Universiteit Leuven, Louvain, Belgium; 4 Université Laval, Quebec City, Quebec, Canada.

Running Title: TcI_{DOM} in chagasic cardiomyopathy and reactivation

Keywords: T. cruzi, lineage, TcI, reactivation, chagasic cardiomyopathy, Chagas disease

1 Corresponding author
Jaime A. Costales
Edificio de Química, PB # 004
Centro de Investigación en Enfermedades Infecciosas
Pontificia Universidad Católica del Ecuador
Avenida 12 de Octubre y Roca, Quito-Ecuador
Abstract

We report an imported case of chronic chagasic cardiomyopathy reactivation in the USA, caused by the TcI_{DOM} subpopulation of Trypanosoma cruzi. Our results dispel the misconception that infection with TcI is benign, and reemphasize the need for increased surveillance, both in endemic areas and in the USA.
Chagas disease, caused by *Trypanosoma cruzi*, is the most important parasitic disease in Latin America, where it affects millions [1]. Moreover, it constitutes an emerging global public health problem, since thousands of *T. cruzi*-infected Latin Americans migrated during the last few decades and now live in North America, Europe, Australia, Japan and other regions [2].

A spectrum of clinical manifestations may result from human infection with *T. cruzi*, ranging from a total absence of symptoms to extremely debilitating and often deadly cardiac or digestive syndromes [3], and *T. cruzi*’s genetic diversity is suspected to play a key role in the clinical outcome. Six major genetic lineages or discrete typing units (DTUs) are currently recognized (named TcI through TcVI), each displaying different biological characteristics [4]. Although no proven associations between *T. cruzi* genetic lineages and the clinical presentations of the disease exist at present, DTUs TcII, V and VI have been frequently reported to be present in serious chronic manifestations in the Southern Cone of South America [4], while TcI has been suggested to be benign, where the observed chronic chagasic cardiomyopathy (CCC) in TcI-infected patients is instead attributable to coinfection with other *T. cruzi* DTUs [5-8]. TcI exhibits high intra-lineage diversity, with specific populations (termed TcI\textsubscript{DOM}) associated with human infection [9], although no direct link with clinical manifestations has been established.

**The patient**

A 43 year old chagasic male from El Salvador was admitted to Massachusetts General Hospital to undergo orthotropic heart transplantation. Immunofluorescence assay performed at the Centers for Disease Control & Prevention (CDC, Atlanta, USA) was positive at >1:256 (cut off =1:32). His cardiac symptoms corresponded to New York Heart Association class IV and he had a biventricular pacemaker/defibrillator for complete heart block. Orthotropic heart transplantation
was successful. Pathologic examination of the explanted heart revealed findings consistent with end-stage CCC, including dilatation with near complete atrophy of the left ventricular wall, endocardial fibrosis, diffuse myocardial fibrosis, and mononuclear infiltrates with some eosinophils and neutrophils. The infiltrate (lymphocytic myocarditis) was composed of lymphocytes (many CD3 T cells, more CD8 than CD4, few CD20 B cells) plus many CD68 macrophages. Amastigotes were not identified on multiple sections examined; however, *T. cruzi* kinetoplast DNA was detected by PCR in frozen tissue submitted to the Parasitic Diseases and Diagnostics Branch of the CDC. Weekly microscopic examination of his blood was performed after transplant, screening for early detection of reactivation disease; this was positive with rare trypomastigotes detected at week six after transplant. He was given nifurtimox for 10 weeks, and developed severe peripheral neuropathy with anorexia. He was switched to benznidazole for 30 days, which he tolerated well. He was monitored monthly for parasitemia for approximately one year after the end of therapy, with no evidence of further infection.

**The study**

Work was approved by MGH’s Internal Review Board. Genotyping directly from patient’s blood samples and parafinized heart explants using a nested PCR-RFLP for the 1f8 flagellar protein and digestion with *Alw* 21I restriction enzyme (Van der Auwera, unpublished) assigned parasites to DTU TcI (Technical Annex 1). Hemoculture six and eight weeks after transplant yielded epimastigotes, which were cloned in solid medium. Cultured parasites and clones were determined to belong to DTU TcI by polymerase chain reaction and restriction fragment length polymorphism (PCR-RFLP) as in [10] (Technical Annex 2). Intra-TcI genotyping was performed with nuclear microsatellites [11] and maxicircle multilocus sequence typing [12]. Intriguingly, microsatellite data indicate a close relationship with TcIDOM, a distinct
genotype within DTU TcI which is common among human cases in Latin America (Figure 2A),
while the maxicircle sequence analysis indicates an origin among wild / non-human isolates for
North and Central America (Figure 2B).

Finally, although patient’s serum reacted against *T. cruzi* antigen in three different
commercial serological tests (Chagatek-Biomerieux, Chagas III-Abbott BiosChile and Chagatest
Recombinante 3.0-Wiener Lab), it did not recognize synthetic peptides derived from the TSSA
antigen specific for DTUs TcII, IV or V/VI described in [13].

**Conclusions**

TcI constitutes the most abundant and widespread *T. cruzi* DTU [4] and is the
predominant (although not the only) DTU in the Amazon region and countries North of it [1].
However, inability to detect TcI in a predominantly TcII endemic region, led to the
misconception that TcI was not a cause of CCC. For our patient, only TcI DTU parasites were
detectable in the clinical samples, hemocultures, and clones. Furthermore, antibodies against
TCII, IV or V/VI specific epitopes were not detectable in serum, suggesting the absence of co-
infection or previous infection with those lineages. Although no TcIII-specific antigen is
currently available, TcIII infects humans only in exceptional cases, and no reports in Central
America exist [4]. Therefore, our data strongly support the conclusion that the observed CCC
was caused exclusively by TcI parasites [5-7]. Previous reports implicating TcI in CCC [14]
have either depended heavily on serological approaches which are currently known to be flawed
[13] or not tested for the presence of coinfection with other lineages [15]. TcI_{DOM} has not
previously been reported from cardiac tissues in CCC cases, as it has been merely detected in
peripheral blood or hemoculture, where coinfection with parasites from other genetic lineages residing in cardiac tissue and being the actual cause of CCC cannot be ruled out.

TcI has considerable intra-DTU diversity [11]; specific genotypes within TcI are associated with human infection [9]. Based on nuclear microsatellite information, the patient was infected by TcIDOM, a genotype associated with many human infections in regions north of the Amazon. Conversely, mitochondrial genotyping suggests a closer relationship with isolates from North and Central America, consistent with local, possibly sylvatic, origin of the infecting strain. Given the proclivity for mitochondrial introgression into TcIDOM [12] we suggest our observation is yet another of such hybridizations, in this case between TcIDOM and a local strain, highlighting the need for control strategies aimed at domiciliary and extrinsic parasite populations as sources of human infection.

Our results demonstrate that TcI, can cause end-stage CCC and dispels the long-held bias that infection with this lineage is benign [5-7]. Considering the wide distribution of TcI (the only T. cruzi DTU ranging from the Southern United States to Argentina and Chile) and the frequency with which TcI strains are associated with human infection [4], there is need for greater surveillance in TcI endemic regions like Central America. Around 22 million people from Chagas endemic countries live in the US, and most of these immigrants come from Mexico (74%) and El Salvador (6.4%) [2], where TcI is known to predominate [4]. Thus, a significant proportion of the estimated 300,000 T. cruzi infections among immigrants in the U.S.A. are predicted to involve the DTU TcI, adding to the growing economic burden of medical care and interventions associated with Chagas’ disease in the U.S.A., including transplantation for end-stage heart disease.
Acknowledgements: We thank Fernanda Latorre, Alejandra Zurita for technical assistance, Gert Van der Auwera for valuable discussion of the data, Mario Grijalva and Michael Miles for critical reading of the manuscript. Work was funded by PUCE.

Jaime Costales is a faculty member at the CIDR/PUCE, Quito-Ecuador. His research interests include the biology of the genetic lineages of *T. cruzi* and the molecular/cellular events in the parasite’s intracellular cycle.

References


Figure 1. Neighbour joining dendrogram based on pairwise inverse allele sharing which shows the relationship between parasite clones isolated from the patient and others from North, Central and South America. Branch colours indicate strain origin and values at important nodes indicate percentage of bootstrap support over 1000 trees. Further details of strains and analytical strategy can be found in [9].
**Figure 2. Maxicircle sequence-based typing of strain isolated from patient.**

Maxicircle sequences for one biological clone were concatenated according to [12], aligned against 70 Tcl strains encompassing Tcl genetic diversity from across North, Central and South America [9] and used to assemble a Maximum-Likelihood topology in PhyML. The best-fit model of nucleotide substitution was selected from 88 models and its significance evaluated according to the Akaike Information Criterion (AIC) in jMODELTEST 1.0. The best fit model selected for this dataset was GTR+G. Bootstrap support for clade topologies was estimated following the generation of 100 pseudo-replicate datasets.
Technical Annex 1. Molecular typing from clinical samples. DNA was extracted from blood samples and parafinized heart explants. Typing was performed using a nested PCR-RFLP strategy for amplification of the 1f8 flagellar protein and digestion with Alw21I (Van der Auwera, et al., unpublished data). H = DNA extracted from parafinized heart explant tissue, B and B2= blood samples taken 10 days apart during reactivation of disease, TcI-TcVI correspond to DTU controls, NC1 corresponds to negative control for the PCR and NC2 for the nested-PCR. Lanes corresponding to Alw21I restriction digest products are labeled with an asterisk (*). Only restriction products are shown for controls.

DNA was extracted from epimastigote hemocultures (HC) and five derived clones (CL1-CL5). Tcl-TcVI correspond to DTU controls, NC corresponds PCR negative control.

Lanes containing restriction products are labeled with an asterisk (*). Only restriction products are shown for controls. A. DNA was analyzed by the PCR-RFLP scheme proposed by Lewis, et al., 2009 [10]: as indicated by the brackets on the right side, fragments from the LSUrDNA, HSP60 and GPI genes were amplified by PCR. GPI and HSP60 products were digested with HhaI and EcoRV restriction enzymes, respectively.

B. Results were confirmed by amplifying strategy the 1f8 flagellar protein and digesting the amplicons with Alw 21I restriction enzyme (Van der Auwera, et al., unpublished data).
COVER SHEET FOR EACH ‘RESEARCH PAPER’ INCLUDED IN A RESEARCH THESIS

Please be aware that one cover sheet must be completed for each ‘Research Paper’ included in a thesis.

1. For a ‘research paper’ already published

   1.1. Where was the work published? PLoS Neglected Tropical Diseases

   1.2. When was the work published? 2014

   1.2.1. If the work was published prior to registration for your research degree, give a brief rationale for its inclusion
   … N/A …

   1.3. Was the work subject to academic peer review? Yes

   1.4. Have you retained the copyright for the work? Yes No

      If yes, please attach evidence of retention.

      If no, or if the work is being included in its published format, please attach evidence of permission from copyright holder (publisher or other author) to include work

2. For a ‘research paper’ prepared for publication but not yet published

   2.1. Where is the work intended to be published?

   2.2. Please list the paper’s authors in the intended authorship order

   2.3. Stage of publication – Not yet submitted / Submitted / Undergoing revision from peer reviewers’ comments / In press

3. For multi-authored work, give full details of your role in the research included in the paper and in the preparation of the paper. (Attach a further sheet if necessary)

   The candidate contributed, laboratory reagents used to generate parasite lysates and infected mouse sera, and technical support during data generation.

NAME IN FULL (Block Capitals) LOUISA ALEXANDRA MESSENGER

STUDENT ID NO: 223021

CANDIDATE’S SIGNATURE ……………………………….. Date 10/2/15 …………..

SUPERVISOR/SENIOR AUTHOR’S SIGNATURE (3 above) ………………………………..
Development of Peptide-Based Lineage-Specific Serology for Chronic Chagas Disease: Geographical and Clinical Distribution of Epitope Recognition

Tapan Bhattacharyya1*, Andrew K. Falconer2, Alejandro O. Luquetti3, Jaime A. Costales4, Mario J. Grijalva4,5, Michael D. Lewis1, Louisa A. Messenger1, Trang T. Tran1, Juan-David Ramirez6, Felipe Guhl7, Hernan J. Carrasco8, Patricio Diosque9, Lineth Garcia10, Sergey V. Litvinov11, Michael A. Miles1

1 Faculty of Infectious and Tropical Diseases, London School of Hygiene and Tropical Medicine, London, United Kingdom, 2 Departamento de Medicina, Universidad del Norte, Barranquilla, Colombia, 3 Laboratorio de Chagas, Hospital das Clínicas, Universidade Federal de Goiás, Goiânia, Goiás, Brazil, 4 Centro de Investigación en Enfermedades Infectiosas, Escuela de Biología, Pontificia Universidad Católica del Ecuador, Quito, Ecuador, 5 Tropical Disease Institute, Department of Biomedical Sciences, Heritage College of Osteopathic Medicine, Ohio University, Athens, Ohio, United States of America, 6 Facultad de Ciencias Naturales y Matemáticas, Universidad del Rosario, Bogotá, Colombia, 7 Centro de Investigaciones en Microbiología y Parasitología Tropical, Universidad de los Andes, Bogotá, Colombia, 8 Universidad Central de Venezuela Instituto de Medicina Tropical, Caracas, Venezuela, 9 Unidad de Epidemiología Molecular, Instituto de Patología Experimental, Consejo Nacional de Investigaciones Científicas y Técnicas (CONICET), Universidad Nacional de Salta, Salta, Argentina, 10 Facultad de Medicina, Universidad Mayor de San Simón, Cochabamba, Bolivia, 11 Aptum Biologics Ltd, Southampton, United Kingdom

Abstract

**Background:** Chagas disease, caused by infection with the protozoan *Trypanosoma cruzi*, remains a serious public health issue in Latin America. Genetically diverse, the species is sub-divided into six lineages, known as TcI–TcVI, which have disparate geographical and ecological distributions. TcII, TcV, and TcVI are associated with severe human disease in the Southern Cone countries, whereas TcI is associated with cardiomyopathy north of the Amazon. *T. cruzi* persists as a chronic infection, with cardiac and/or gastrointestinal symptoms developing years or decades after initial infection. Identifying an individual’s history of infection of individuals, geographical and clinical associations of *T. cruzi* lineage infection directly by genotyping of the parasite is complicated by the low parasitaemia and sequestration in the host tissues.

**Methodology/Principal Findings:** We have applied here serology against lineage-specific epitopes of the *T. cruzi* surface antigen TSSA, as an indirect approach to allow identification of infecting lineage. Chagasic sera from chronic patients from a range of endemic countries were tested by ELISA against synthetic peptides representing lineage-specific TSSA epitopes bound to avidin-labelled ELISA plates via a biotin-labelled polyethylene glycol-glycine spacer to increase rotation and ensure each amino acid side chain could freely interact with their antibodies. 79/113 (70%) of samples from Brazil, Bolivia, and Argentina recognised the TSSA epitope common to lineages TcII/TcV/TcVI. Comparison with clinical information showed that a higher proportion of Brazilian TSSApep-II/IV/VI responders had ECG abnormalities than non-responders (38% vs 17%; p<0.0001). Among northern chagasic sera 8/42 (20%) from Ecuador reacted with this peptide; 1/12 Venezuelan and 1/34 Colombian samples reacted with TSSApep-IV. In addition, a proposed TcI-specific epitope, described elsewhere, was demonstrated here to be highly conserved across lineages and therefore not applicable to lineage-specific serology.

**Conclusions/Significance:** These results demonstrate the considerable potential for synthetic peptide serology to investigate the infection history of individuals, geographical and clinical associations of *T. cruzi* lineages.


Editor: Alain Debrabant, US Food and Drug Administration, United States of America
Received February 7, 2014; Accepted April 9, 2014; Published May 22, 2014

Copyright: © 2014 Bhattacharyya et al. This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Funding: This work was supported by the European Commission Framework Programme Project “Comparative epidemiology of genetic lineages of *Trypanosoma cruzi*” ChagasEpNet, Contract No. 223034; http://cordis.europa.eu/fp7/home_en.html. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The epitope prediction was performed as a research collaboration with Sergey V. Litvinov of the company Aptum Biologics Ltd, without payment; none of the authors have any financial, non-financial, professional or personal conflicting interests; this collaboration does not alter our adherence to all PLOS NTDs policies on shared data and materials.

* E-mail: Tapan.Bhattacharyya@lshhtm.ac.uk

Introduction

Chagas disease (South American trypanosomiasis) is still considered to be the most important parasitic disease in Latin America, despite notable success with control of household infestation by the triatomine insect vectors. Up to 8 million people are estimated to be chronically infected with the causative agent *Trypanosoma cruzi*, of whom at least 30% are likely to develop chagasic cardiomyopathy, in some cases with megasystoles of the intestinal tract [1,2]. Vector borne transmission is usually by
contamination of mucous membranes or abraded skin with \textit{T. cruzi} infected triatomine faeces and sporadic oral outbreaks occur due to triatomine contamination of food [3]. Infection can also be propagated by congenital transmission and blood or organ donation, and this may arise among migrant populations far beyond the endemic regions in Latin America [4].

The species \textit{T. cruzi} is remarkably diverse genetically and is currently described as comprising six distinct lineages or discrete typing units (DTUs, TcI-TcVI) [5]. The six lineages have complex disparate but partially overlapping geographical and ecological distributions and are circumstantially associated with different epidemiological features [6,7]. TcI is the principal agent North of the Amazon, in association with chagasic heart disease but where megasymphyses are considered to be rare. TcII is one of three principal agents of Chagas disease in the Southern Cone region of South America, where chagasic cardiomyopathy, megaesophagus and megacolon are found. TcIII is seldom isolated from humans but is widely distributed with the natural armadillo host \textit{Dasypus novemcinctus}. TcIV is a sporadic secondary agent of Chagas disease in Venezuela [8]. TcV and TcVI, like TcII, are also agents of Chagas in the Southern Cone region, and are known to be relatively recent hybrids of TcII and TcIII [9,9].

Parasitological diagnosis in the acute phase of \textit{T. cruzi} infection is by microscopy of fresh blood films, thin blood films, thick blood films or by haematoцит centrifugation and examination of the buffy coat, the latter being recommended particularly for congenital cases. In the chronic phase recovery of live organisms may be attempted by multiple blood cultures or xenodiagnosis with colony bred triatomine bugs but with limited sensitivities, or parasite DNA may be detectable by amplification.

Serological diagnosis of \textit{T. cruzi} infection is usually performed by either indirect immunofluorescence (IFAT) or indirect haemagglutination (IHA) or enzyme-linked immunosorbent assay (ELISA), giving >94% sensitivity and specificity [2]. There are several commercially available diagnostic kits, including rapid lateral flow tests but sensitivities may not be equivalent, particularly when they are used in regions where non-homologous genetic lineages of \textit{T. cruzi} are prevalent [8–10]. These serological methods give no information on the genetic lineage or lineages that a patient carries, and are not designed for that purpose.

A key objective of Chagas research therefore remains to follow up in detail the circumstantial evidence of a relationship between infecting \textit{T. cruzi} lineage and the clinical outcome [6,7,11]. However, such analysis is complex and vulnerable to multiple confounders, including diversity of host susceptibility. Even if \textit{T. cruzi} isolates can be recovered from the infected blood by parasitological diagnosis or if DNA can be amplified from blood, genotyping methods [12,13] do not provide an entire profile of the infecting lineages in an individual patient, because distinct \textit{T. cruzi} lineages may be sequestered in the tissues [14]. An approach to overcoming this limitation is to identify infecting \textit{T. cruzi} lineage in a more indirect way. One strategy to achieve this is by serological detection of antibodies that are produced in response to lineage-specific antigens.

Di Noia et al [15] described the trypomastigote small surface antigen (TSSA), encoded by a member of the \textit{TAMUCHI} mucin gene family, expressed on the mammalian bloodstream trypomastigote stage of the \textit{T. cruzi} life cycle. The authors reported that TSSA is dimorphic in sequence, with TSSA-I being present in TcI, and TSSA-II found in TcII-TcVI. On the basis of this finding the authors pioneered lineage-specific serology for Chagas disease through application of a TSSA-II recombinant antigen to serology with patients from the Southern Cone region of South America. Chagasic patients were only TSSA-II seropositive, which led to the suggestion that TcI could be benign. However, this suggestion was in conflict with the geographical predominance of TcI North of the Amazon and the acute and chronic clinical presentations of known TcI infections [16,17]. In subsequent publications \textit{E. coli}-produced recombinant TSSA proteins have been used more widely for serology with humans and animals [18–23].

We have previously analysed TSSA diversity among a panel of \textit{T. cruzi} isolates representing a broad geographical and ecological range of lineages TcI-TcVI [24]. We found a greater lineage-specific diversity than had previously been described. Lineages TcII, TcV, and TcVI were shown to share a common TSSA sequence. However, in both of the hybrid lineages TcV and TcVI we found that two TSSA alleles were present at an heterozygous locus within the polymorphic epitope: one haplotype was shared with TcII and in the second haplotype a Thr was replaced by Ala at position 44 of the protein. Lineage-specific TSSA sites were also found in TcIII and TcIV strains [24]. Gánpéa et al [25] suggested a functional significance for this diversity in that the TcI/TcV/TcVI form of TSSA, but not the TcI form, has the property of binding surface receptor(s) and inducing signalling pathways in host cells prior to parasite internalisation.

Recently, Mendes et al [26] used a bioinformatic analysis of the reference genome of the TcVI strain CL Brener [27] to identify candidate peptides for differential screening with sera from mice experimentally infected with single, known \textit{T. cruzi} lineages. A resultant peptide, derived from a putative RNA-binding protein, was reported to be applicable for TcI serology [26].

Here, we have used our expanded knowledge of the range of TSSA diversity to design and synthesise lineage-specific peptides. We assess the capacity of these peptides to provide antigens for lineage-specific serology by ELISA and thus reveal which lineages have infected individual patients during their lifetime. Furthermore, we examine the geographical and clinical distribution of recognition of the synthetic peptide epitopes. In addition, we also investigate the diversity of the gene coding for the peptide described [26] as applicable for TcI-specific serology.
Materials and Methods

Ethics statement

Human sera were collected as part of routine diagnostic examination, with local institutional ethical approvals, and in accord with EC ethical standards established as part of the ChagasEpiNet international collaboration. All human sera were anonymised and coded by letters and numbers that did not reveal patient identities. Production of mouse sera adhered to the European 3Rs policy of Refinement, Reduction and Replacement (99/167/EG; Council decision of 25/1/99), took place in authorised animal facilities by licensed staff in agreement with the European Directive 86/609/EEC, and with review and approvals under UK Home office regulations [Animals (Scientific Procedures) Act 1986; project licence number 70/6997 to the London School of Hygiene and Tropical Medicine].

Mouse and human sera

Mouse sera were from mice previously inoculated intraperitoneally with 10^6 organisms from stationary phase cultures representing the lineages. Sera were separated from whole mouse blood by allowing clotting at room temperature, overnight storage at 4°C, centrifugation at 12,000xg for 10 mins and removal of the supernatant serum. Serum samples were stored 1:1 with glycerol at −20°C.

Human sera were from chronic cases of Chagas disease, confirmed by a combination of parasitological and serological diagnosis. As shown in Table 1, 113 samples were from the Southern Cone countries, Brazil, Bolivia and Argentina, and 66 samples were from countries North of the Amazon, Colombia, Ecuador, Venezuela, where TcI has been considered to predominate. Brazilian sera were from patients who had a positive parasitological diagnosis at the time of serum collection, together with a full clinical history, their geographical origin, age and sex. Institutes providing sera were: Hospital das Clinicas, Goiaia, Brazil; Universidad Mayor de San Simon, Cochabamba, Bolivia; Universidad Nacional de Salta, Argentina; Universidad Central de Venezuela, Caracas, Venezuela; Universidad de los Andes, Bogota, Colombia; Pontificia Universidad Catolica del Ecuador, Quito, Ecuador. Endemic healthy controls were 17 sera from Colombia that were serologically negative to T. cruzi lysate.

Synthesis of lineage-specific peptides

The synthetic peptides were prepared with an amino terminal biotin molecule linked via a polyethylene glycol-glycine spacer so that they could be bound to avidin-coated ELISA plates. Importantly, this method increased their rotation and ensured that each amino acid side chain could freely interact with antibodies, as opposed to being adsorbed onto the solid phase where some amino acid side chains would be unavailable, as discussed previously [28].

Design of the peptides was based on the T. cruzi TSSA lineage-specific amino acid sequences previously described [24]; chimeric peptides comprised by TSSA-I and TSSA-II sequences were also designed and synthesised (Results; Figure 1). Synthetic peptides were prepared at the 20 μM scale on 100–200 mesh-size Fmoc-Cys(Trt) Wang resin (0.5 mmol/g) [856006; Novabiochem, UK] using a Zinsser Analytic SMPS 350 (Zinsser Analytic, UK) or Advanced Chemtech Apex 396 (Advanced Chemtech, USA) robotic multiple peptide synthesizer. Aspartamide formation of aspartic acid residues was reduced by the use of OMpe-protected Fmoc-Asp(OMpe)-OH (852104; Novabiochem, UK). The coupling steps were performed using 0.5 M Fmoc-protected amino acids diluted in 6.76% (wt/vol) 1-hydroxybenzotriazole (HOBt)/dimethylformamide (DMF) (Activote, UK/Rathburn Chemicals Ltd., UK) activated using 0.5 M N,N,N’,N’-tetramethyl-O-(1H-benzotriazol-1-yl)-uronium hexafluorophosphate (HBTU) (851006; Novabiochem, UK) with 1M N,N-disopropyl-ethylamine (DIPEA) (Rathburn Chemicals Ltd., UK), while the deprotection steps were performed using 20% (vol/vol) piperidine/DMF (Rathburn Chemicals Ltd., UK). The carboxyl- and amino- groups flanking the core epitope sequences contained additional glycine (G) residues to increase rotation (high dihedral (ψ against φ) angles) of their carbonyl-terminal cysteine (C) residue and their amino-terminal spacer and molecular label. Their amino termini were labelled via a polyethylene glycol (PEG) spacer (Fig 1A) through sequential couplings with 0.5 M Fmoc-PEG6-COOH (13 atoms or 20 atoms) (851034 or 851035; Novabiochem, UK) followed by 0.5 M biotin (B4501; Sigma Aldrich, UK) using the more efficient coupling agent, 0.5 M N,N,N’,N’-tetramethyl-O-(7-azabenzotriazol-1-yl)-uronium hexafluorophosphate (HATU) (851013; Novabiochem, UK) containing 1 M DIPEA. The final peptides were each washed 5 times with dichloromethane and then methanol (Rathburn Chemicals Ltd., UK) before being dried in a freeze-drier (Edwards, UK). Peptide cleavage was performed by reaction for 3–4 hours using 1% (wt/vol) phenol, 2% H2O, 2.5% (vol/vol) trisopropyl silane (233781; Sigma Aldrich, UK) and 2% (vol/vol) 2,2’-ethylendioxy diethanethiol (3,6-dioxa-1,8-octanedithiol (DODT)) (465178; Sigma Aldrich, UK) in trifluoroacetic acid (Rathburn Chemicals Ltd., UK) [29]. The cleaved peptides were then precipitated in cold (0°C) peroxide-free diethyl ether (Rathburn Chemicals Ltd., UK), centrifuged at 2,000xg; the supernatants were discarded and the precipitation and centrifugation steps were repeated twice. The peptides were then dried under a stream of anhydrous argon gas (BOC, UK) before being stored at −80°C.

For purification each peptide was subsequently dissolved in 500 μl of 2,2,2-trifluoroethanol (T63002; Sigma Aldrich, USA), made to 5.5 ml with 5% (vol/vol) fur UV grade acetonitrile (Rathburn Chemicals Ltd., UK) in H2O containing 0.1% (vol/vol) trifluoroacetic acid, and then subjected to reversed-phase HPLC using a 5 ml injection loop, a 5–95% acetonitrile gradient run at 9 ml/min over 30 min through an ACE C18-300 Å 250×2.1 mm preparative column (ACE-231-2520: Advanced Chromatography Technologies, UK) in a Beckman Gold preparative HPLC system (Beckman, USA). The main peaks, detected at a wavelength of 215 nm, were collected and freeze-dried before storage at −80°C.

Additionally, peptides TSSApep-II/V/VI, chimera TSSApep-I/-II, and MenTcI were also synthesised commercially (Genosphere Biotecnologies, Paris, France).

Purified peptides were prepared as 1 mg/ml stock solutions in PBS and the addition of biotin in the last coupling was assessed by ELISA. For this assay, 10 μg/ml of each peptide was prepared in 1× carbonate-bicarbonate coating buffer (15 mM Na2CO3, 34 mM NaHCO3, pH 9.6) and added at 50 μl/well to 96-well ELISA plates (735–0463: Immulon 4HBX, VWR, UK). Plates were covered with an adhesive sheet and incubated overnight at 4°C. Following three washes with PBS containing 0.05% (vol/vol) Tween 20 (P7949; Sigma Aldrich, UK) (PBS/T), 200 μl/well blocking buffer (PBS/2% skimmed milk powder (Premier International Foods, Spalding, UK) was added and incubated at 37°C for 2 hrs. Following three washes, a 1:2000 dilution of peroxidase-labeled streptavidin (S2438; Sigma Aldrich, UK) in PBS/T containing 2% skimmed milk powder (PBS/T/M) was added at 50 μl/well, and incubated at 37°C for 1 hour. After washing six
Table 1. Geographical distribution of antibody responses to lineage-specific synthetic peptides, as determined by ELISA.

<table>
<thead>
<tr>
<th>TSSA peptide reaction</th>
<th>Chimera</th>
<th>Non-specifica</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lineage-specific</td>
<td></td>
<td></td>
</tr>
<tr>
<td>n</td>
<td>I</td>
<td>II/VI</td>
</tr>
<tr>
<td>Brazil</td>
<td>98b</td>
<td>1b</td>
</tr>
<tr>
<td>Bolivia</td>
<td>10a</td>
<td>0</td>
</tr>
<tr>
<td>Argentina</td>
<td>5</td>
<td>1d</td>
</tr>
<tr>
<td>Colombia*</td>
<td>34</td>
<td>0</td>
</tr>
<tr>
<td>Ecuador</td>
<td>20a</td>
<td>0</td>
</tr>
<tr>
<td>Venezuela</td>
<td>12</td>
<td>0</td>
</tr>
<tr>
<td>EHC (Brazil)</td>
<td>7</td>
<td>0</td>
</tr>
<tr>
<td>TOTAL</td>
<td>186</td>
<td>2</td>
</tr>
</tbody>
</table>

EHC = Endemic healthy controls (* a further 17 Colombian sera that were serologically negative with the lysate were included in the peptide ELISAs as additional controls); ND = not determined.

a these 98 comprised 1 sample from each of 90 patients, plus 2 paired samples from each of 4 patients. All eight paired samples reacted with TSSApep-II/V/VI, and are included within the 67 Brazilian reactors to this peptide. 1 set of these pairs also reacted with TSSApep-V/VI.
b same sample, which did not react with TSSApep-II/V/VI, TSSApep-V/VI or chimeras.
c these 9 samples also reacted with chimera TSSA-II/-I peptide.
d same sample, which did not react with TSSApep-II/V/VI, TSSApep-V/VI or chimeras.
e same sample.
f same sample, which did not react with TSSApep-I, TSSApep-II/V/VI, TSSApep-V/VI or chimera TSSApep-II/IV.
g non-specific binding; see text.
h in each case the same sample reacted with TSSApep-III and TSSApep-IV.
doi:10.1371/journal.pntd.0002892.t001
times with PBS/T, 50 mM phosphate/citrate buffer (pH 5.0) containing 2 mM o-phenylenediamine HCl (P1526; Sigma Aldrich, UK) and 0.005% (vol/vol) H₂O₂ (216763; Sigma Aldrich, UK) was added at 50 µl/well and the plates were incubated in the dark at room temperature for 10 minutes. The substrate reactions were then stopped by the addition of 2M H₂SO₄ (25 µl/well) and the absorbance values were determined at a wavelength of 490 nm (MRX, Dynatech, USA).

Production of whole-cell lysate antigen

*T. cruzi* was cultured as previously described [30]. For production of lysate antigen, mid-to-late log phase cultures of a TcII strain (IINF/PY/00/Chaco23cl4) of *T. cruzi* were centrifuged at 13000 rpm for 1 min, and the supernatant used for antigen in ELISA. Protein concentration was determined using the BCA Protein Assay kit (PN23227: Fisher Scientific, UK).

Lineage-specific peptide ELISA

Immulon 4HBX 96-well flat bottomed ELISA plates were coated with 1 µg/100 µl/well of avidin (A9275; Sigma, UK) diluted in 1× carbonate-bicarbonate coating buffer for binding to lineage-specific peptide, and in separate wells coating was with TcII *T. cruzi* lysate at 0.2 µg/100 µl/well to act as a serologically positive control for each sample. Plates were covered with an adhesive sheet and incubated overnight at 4°C. The following day, unbound avidin and lysate were removed, the plate washed three times with wash buffer PBS/T, then wells were blocked with 200 µl blocking buffer PBS/T/M at 37°C for 2 hrs. Following three washes, 1 µg/100 µl/well TSSA lineage-specific peptide in PBS/T/M was incubated with the avidin-coated wells at 37°C for 1 hr. Following three washes, 100 µl/well of 1:200 dilution of serum in PBS/T/M was added and incubated at 37°C for 1 hr. Following six washes, 100 µl/well of donkey anti-human IgG (H+L)-HRP (709-035–149; Jackson Immunoresearch, Pennsylvania, USA), diluted 1:5000 in PBS/T/M was added, and incubated at 37°C for 1 hr. Following six washes, plates were developed and read as described above, except that the volumes were 100 µl for substrate and 50 µl for 2M H₂SO₄. Replica plates were run in duplicate simultaneously.

Statistical analysis

Cut-off values for ELISAs with human sera and peptides were calculated from the mean plus 3 standard deviations compared to the endemic healthy controls from Goiânia, Brazil. Statistical analysis (2-tailed unpaired t-test) on the Brazilian TSSApep-II/V/VI seropositives and non-responders was performed using GraphPad Prism version 4.02 for Windows (GraphPad Software, San Diego, California, USA).

Analysing diversity of the putative TcI epitope

We designed PCR primers MenTcI FOR (5’ ATGCCCA-CAATCGAAACCAAG 3’) and MenTcI REV (5’ TGACAA-CAAACGTGTTGGCTG 3’) (synthesised by Eurofins MWG Operon, Germany) to amplify the whole open reading frame (ORF) of the putative RNA-binding protein (Tc00.104703511837.129) which was described as containing an epitope and corresponding peptide applicable for TcI serology [26]. *T. cruzi* strains, from which genomic DNA was used as amplification template, are listed in Table 2. Amplification reactions were performed in a total volume of 20 µl, and comprised of 1×NH₄ reaction buffer supplemented with 1.5 mM MgCl₂ (Bioline, UK), 200 mM dNTPs (New England Biolabs, UK), 10 pmol of each primer, and 1 U BioTaq DNA polymerase (Bioline). Amplification conditions were: 1 cycle of 94°C, 3 mins; 25 cycles of 94°C for 30 secs, 55°C for 30 secs, 72°C for 30 secs; 1 cycle of 72°C for 10 mins. Five microliters of the PCR reaction were analysed by electrophoresis on 1.5% agarose gels (Bioline); amplification products were purified from the remaining reaction by precipitation with an equal volume of isopropanol at room temperature, followed by washing with 70% EtOH, air-drying and resuspension in ddH₂O. Bi-directional DNA sequencing, using each PCR primer separately at 3.2 pmol, was achieved using a BigDye

---

**Figure 1. TSSA provides potential epitopes that are *T. cruzi* lineage-specific.** [A] Components of the peptides synthesised: N-terminal biotinylation; PEG spacer; Gly; the lineage-specific sequence; C-terminal Cys. [B] Amino acid sequences of the synthetic peptides (TSSApep-), with polymorphic residues underlined; for the two chimeric peptides the TSSA-II residues are shown in bold.

doi:10.1371/journal.pntd.0002892.g001

---

**Table 2**

<table>
<thead>
<tr>
<th>PEPTIDE</th>
<th>AMINO ACID SEQUENCE</th>
</tr>
</thead>
<tbody>
<tr>
<td>TSSApep-I</td>
<td>GTDKKTAAGGTPSPSG</td>
</tr>
<tr>
<td>TSSApep-II/V/VI</td>
<td>GTENKPATGEAPSQPG</td>
</tr>
<tr>
<td>TSSApep-III</td>
<td>GTEKKAAGEAPSPSG</td>
</tr>
<tr>
<td>TSSApep-IV</td>
<td>GTDKKTAAGEAPSPSG</td>
</tr>
<tr>
<td>TSSApep-V/VI</td>
<td>GTENKPAAGEAPSPQPG</td>
</tr>
<tr>
<td>Chimera TSSApep-I/-II</td>
<td>GTDKKTAGETPSPSG</td>
</tr>
<tr>
<td>Chimera TSSApep-III/I</td>
<td>GTENKPATGETPSPSG</td>
</tr>
</tbody>
</table>
Terminator v3.1 RR-100 kit (Applied Biosystems, UK) according to standard protocols. Sequence alignment was performed using BioEdit software [31]. In parallel, the coding region of the TSSA gene containing lineage-specific sequences was also sequenced, as described previously [24], to confirm lineage identity.

### Linear B-epitope profiling

Computer analysis of the TSSA-I and the TSSA-II/V/VI common epitope was performed using EpiQuest-B software (v 2.1.17, Matrix B7.1) from Aptum Biologics Ltd (Southampton, Hampshire, UK). The algorithm of the program allows prediction of potential linear B-epitopes and their immunogenicity. The data were used in graphical format.

### Accession numbers

Nucleotide sequences derived in this manuscript are available under GenBank accession numbers KJ395471 - KJ395493.

### Results

**TSSA provides potential epitopes specific for each T. cruzi lineage**

The structures and sequences of the peptides synthesised, indicating the lineage-specific amino acids, are shown in Figure 1, as based on the comparisons of diversity previously described [24]. In addition to the peptides representing single lineages we synthesised two chimeric peptides, one with TSSA-I residues at the N terminus and TSSA-II residues at the C terminus, and the second with TSSA-II at the N terminus and TSSA-I at the C terminus (Figure 1B).

Consistent with the known extensive genomic divergence between TcI and TcII, eight residues differed between their TSSA potential epitopes. Five and six residues separated TSSA-II from TSSA-III and two residues separated TSSA-I from TSSA-IV, in accord with their somewhat greater affinity with TcI. A single residue differed between the TSSA-II haplotype shared by TcII, TcV and TcVI and the second haplotype present at the heterozygous locus in the hybrids TcV and TcVI.

### Synthetic peptides are recognised by serum antibodies

Sera from mice experimentally infected with biological clones of TcII, TcV and TcVI strains recognised TSSApep-II/V/VI in serology by ELISA, and sera from TcIII and TcIV murine infections reacted with the corresponding TSSA peptides (Bhat-charyya et al, in preparation), encouraging the evaluation described here of the diagnostic potential of all the synthetic peptides with sera from patients with chronic Chagas disease.

Figure 2 shows examples of ELISA plates with T. cruzi lysate and lineage-specific synthetic peptides as antigens. Sera from normal healthy endemic controls did not react with the T. cruzi lysate or with any of the synthetic peptides. Without exception all
sera from patients with chronic Chagas disease recognised the T. cruzi TcII lysate antigen preparation. Figure 2 also provides examples of sera from Brazil, Argentina and Ecuador that recognised TSSApep-II/V/VI, indicative of infection with TcII, TcV or TcVI. A positive result for the epitope derived from the TcV/VI specific haplotype indicates definite infection with TcV or TcVI. Some of these sera (e.g. B6 & B10) reacted with both TSSApep-II/V/VI and TSSApep-V/VI representing the haplotype restricted to TcV and TcVI, indicating infection with a hybrid strain, possibly in conjunction with a TcII infection. Recognition of the TcV and TcVI restricted epitope was never seen in the absence of recognition of the TSSApep-II/V/VI.

A Venezuelan serum (V4 in Figure 2) recognised TSSApep-IV, consistent with the known presence of TcIV as a secondary agent of Chagas disease in Venezuela [16]. However, this serum also bound to TSSApep-III, which only differs by 2 of 16 residues.

An antibody response to TSSApep-I was exceptional, only two sera were reactive from the entire set of samples (Table 1) of which one weak reactor (Argentina A3) is shown in Figure 2. Chimeras TSSApep-I/-II and chimera TSSApep-II/-I were designed to determine whether the antigenic epitope resided at the N or C terminus of the peptides. Chimera TSSApep-I/-II was recognised by 71/83 TSSApep-II/V/VI reactive sera, as demonstrated for example by Argentine patients A1, A2 and A4 (Figure 2). In comparison, only 11/83 recognised the chimera TSSApep-I/-II, indicative that, although not precisely mapped, the dominant region of the epitope lies towards the N-terminus of the peptide and that in some patients the N terminus is adequate to provide a detectable epitope. A single TSSApep-IV/TSSApep-III positive serum also recognised the chimera TSSApep-I/-II peptide.

Four of 186 samples responded to all wells containing peptides; these were demonstrated to bind non-specifically to avidin in the absence of peptide, but not to cross react with milk proteins (data not shown).

Rare recognition of the TSSA-II/V/VI common peptide in northern South America

The 186 sera from patients with chronic Chagas disease spanned a geographical range from Argentina to Venezuela. Three Southern Cone countries were included, where TcII, TcV and TcVI have been reported to be endemic, and three countries from northern South America, where TcI is considered to predominate. A summary of the geographical distribution of the antibody responses to all the lineage-specific synthetic peptides is shown in Table 1. Of the sera recognizing TSSApep-II/V/VI, 79 out of 83 were from the Southern Cone countries and four were from Ecuador. Of these 83 sera, 13 sera also recognised TSSApep-V/VI, 12 from Southern Cone countries and one of the four sera from Ecuador, indicating presence of TcV or TcVI, possibly with TcII co-infection. Independently of the lineage-specific peptides, we also examined the response to two different chimera peptides, each comprising different combinations of sequences from TSSApep-I and TSSApep-IV. Of the Bolivian, Ecuadorian and Argentine sera which reacted with TSSApep-II/V/VI, all reacted with chimera TSSApep-I/-II, but only two samples (Ecuadorean) also reacted with chimera TSSApep-I/-II. In the case of Brazilian samples, of the 67 that reacted with TSSApep-II/V/VI, 55 reacted with chimera TSSApep-I/-I, and of these 55, 9 also reacted with chimera TSSApep-I/-II. Only one sample (Venezuelan) reacted with chimera TSSApep-I/-II but not with TSSApep-II/V/VI or chimera TSSApep-I/-I. TSSApep-I failed to detect antibodies, regardless of origin of the chagasic sera, with the exception of two sera, one each from Brazil and Argentina. Four sera recognised both TSSApep-IV and TSSApep-III, consistent with cross-reaction due to the close similarity between these epitopes.

The country by country distribution of antibody recognition of the peptides is given in Table 1. ELISA cut-offs and absorbance values for each lineage-specific peptide are shown in Figure 3. Each data point represents the mean A490 readout of duplicate assays of the serum sample with the lineage specific peptides. In Figure 3, the samples giving the highest reading for TSSApep-III from Colombia and Venezuela are the same samples that recognised TSSApep-IV.

Antibodies to the TcII/TcV/TcVI peptide are more frequent among symptomatic Brazilian patients

60/63 of the Brazilian patients with chronic Chagas disease who were seropositive against TSSApep-II/V/VI had detailed clinical evaluation, and of these 60 patients, 23 (38%) had ECG abnormalities typical of Chagas disease. 23/28 patients seronegative for TSSApep-II/V/VI also had detailed clinical evaluation, but in contrast only 4 of these latter, different 23 patients had such ECG abnormalities (p<0.0001).

Novel bioinformatic algorithms predict highly antigenic residues

The sequences coding for the TSSA proteins containing the TSSApep-I and TSSApep-II/V/VI epitopes were subjected to a...
novel bioinformatic analysis using EpiQuest-B program that builds the immunogenicity profile for linear protein sequences and predicts the location and potential immunogenicity of the linear B-cell epitopes (Litvinov et al, in preparation) in order to give an antigenicity score for the polymorphic region. The algorithm predicted high scores within the TSSApep-II/V/VI epitope region, but much lower for TSSApep-I, as shown in Figure 4.

The different antigenicity scores indicate that the N-terminal TEN in TSSApep-II/V/VI is the dominant epitope, explaining the frequent recognition of chimera TSSApep-II/-I (Figure 1) despite the higher scoring C-terminal GEAPS, which provides a secondary site of recognition for some (Brazilian) sera that respond to chimera TSSApep-I/-II (Figure 1). Neither the TEN nor GEAPS is present in TSSApep-I, explaining the rare recognition of this epitope. The GEAPS, which is also present in TSSApep-III and TSSApep-IV, gave much lower antigenicity scores in these peptides, in the absence of the upstream TEN in these sequences (data not shown).

Comparative diversity of the putative Tcl-applicable peptide reveals high conservation across lineages

Mendes et al [26] used heterozygous loci in the CL Brener genome to identify candidate lineage-specific epitopes. One conserved and three potentially lineage-specific peptides were synthesised, described as deriving from hypothetical protein Tc00.1047053510421.310 (conserved epitope peptide), putative RNA-binding protein Tc00.1047053511837.129 (for Tcl serology), putative ADP-ribosylation factor GTPase activating protein Tc00.1047053511589.70 (for Tcl II), and putative DNA-directed RNA polymerase III subunit Tc00.1047053510359.320 (for Tcl VI) that displayed some discriminatory power in ELISAs and affinity-ELISAs based on differential range of absorbance values. The putative Tcl epitope was described as restricted to Tcl and Tcl VI and applicable to serological identification of a specific response to Tcl.

We examined the diversity in the coding sequence for this epitope, using a panel of T. cruzi strains across the lineages (Table 2). A single amplicon of 381bp was produced by PCR from all strains using primers MenTc1 FOR and MenTc1 REV. Examples from Tcl and Tcl II strains are shown in Figure 5.

However, in comparative sequencing across isolates representing the lineages we found this epitope to be highly conserved (Table 2). This epitope had the same amino acid sequence across all the strains and lineages analysed here with the exception of strains Y and CBB (Tcl II) and Para 6 (Tcl V). In strains Y and CBB, a heterozygous nucleotide (G/C) residue led to the presence of two predicted amino acids, Ser or Thr.

The region homologous to the reported Tcl-applicable epitope, which was described as specific to a Tcl II strain, but given only as amino acid sequence [26], was subject to BLAST against NCBI and TriTrypDB databases. There were very low stringency homologies returned by TriTryp BLAST to various T. cruzi proteins (mainly around the PPP tripeptide), none of which was described as RNA-binding proteins. On NCBI BLAST, highest-scoring matches were to various bacteria and fungi, none to trypanosomes. Furthermore, none of 55 sera from northern countries of South America, where Tcl is highly endemic, bound to the Tcl synthetic peptide reported in Mendes et al in our ELISA assays (data not shown).

Discussion

Kong et al [32] developed lineage-specific serology for the protozoan parasite Toxoplasma gondii, which is difficult to isolate from chronically infected patients, and most isolates of which are classified into clonal lineages type I, II, or III. Serology with synthetic peptides based on diversity within the dense granule proteins GRA6 and GRA7 was able to distinguish type II from non-type II infections in humans. Using discriminatory serology, it was possible to demonstrate that the T. gondii lineages had different
continental distributions [33], and that adult offspring of type I-infected mothers had a significantly increased risk for the development of psychoses [34]. Here we have used detailed comparative analysis of the genetic diversity of the T. cruzi TSSA gene encoding the protein core of the mucin TSSA, to design synthetic peptides for lineage-specific serology of T. cruzi infection history. These epitopes were presented on an avidin-coated solid phase via an amino terminal biotin-label linked to a polyethylene glycol-glycine spacer to increase rotation and ensure that each amino acid side chain could freely interact with antibodies. We synthesised and tested these lineage-specific peptides in ELISA with 186 human sera from six countries, three in the Southern Cone region of South America and three in northern South America. We have selected these countries because typing of T. cruzi isolates with multilocus enzyme electrophoresis (MLEE) [6], multilocus sequence typing (MLST) [35] and multilocus microsatellite typing (MLMT) [17] has repeatedly indicated the predominance of TcII, TcV and TcVI as the agents of Chagas disease in Southern Cone countries yet the contrasting high prevalence of TcI in patients North of the Amazon [7,11]. Nevertheless, there have been some reports that TcII, as identified by genotyping, can be found among isolates from humans and domestic triatomine bugs in northern endemic regions, for example in Colomba and Guatemala [36–38].

Lineage-specific serology is therefore of special interest for T. cruzi, because of the disparate geographical distributions of both the T. cruzi lineages and clinical manifestations of chronic Chagas disease. Thus as long ago as 1981, Miles et al proposed that the presence of chagasic cardiomyopathy with megaoesophagus and megacolon in Southern Cone countries, yet apparent absence of associated megasymptoms from Venezuela, may be related to the comparative predominance of Tc as the agent of Chagas disease in northern South America [6]. Nevertheless the evidence of a link between infecting lineage of T. cruzi and prognosis of chronic Chagas disease remains circumstantial. As with T. gondii, this is partly due to the difficulty of isolating and genotyping T. cruzi from chronic chagasic patients. Blood culture and xenodiagnosis have limited sensitivity and may be selective for faster growing biological clones. Furthermore, even if parasites or DNA can be recovered from chronically infected patients, the resultant T. cruzi isolates may not be representative of the genetic diversity in the patient, because T. cruzi replicates intracellularly and lineage genotypes may be sequestered in the tissues but not recoverable from the circulating blood [14]. Serology with lineage-specific antigens provides a means of profiling an individual’s history of T. cruzi infection, to overcome inaccessibility of the parasite to direct genotyping during chronic infections.

TSSA provides a good candidate for development of synthetic peptide-based, lineage-specific serology, because no TSSA homologue beyond the species T. cruzi has been detected by genomic comparisons, and a lineage-specific candidate epitope can be represented by a single synthetic peptide. Thus such peptides are unlikely to generate false positive ELISA results with sera from endemic healthy controls or from patients with other infectious or autoimmune diseases. In the multiple ELISAs performed here none of the healthy controls recognised any of the synthetic peptides, and all were also serologically negative with the T. cruzi lysate (Figure 2). However, sera from four of the chagasic patients bound non-specifically to plates coated with avidin alone and thus

Figure 4. Computer-predicted antigenicity score is much higher for TSSA-II/V/VI sequence than TSSA-I. Polymorphic sequences of [A] TSSA-I and [B] TSSA-II/V/VI showing regions of high antigenicity in red, and low antigenicity in blue. A few amino acid replacements in TSSA-I lead to disappearance of the immunogenic epitope that is present in TSSA-II/V/VI sequence. doi:10.1371/journal.pntd.0002892.g004

Figure 5. PCR amplification of the ORF containing the reported TcI-applicable epitope. Only the amplicon of predicted size (381 bp) was amplified by the reaction conditions. –ve = no template DNA control. Mk = Hyperladder IV (Bioline, UK). doi:10.1371/journal.pntd.0002892.g005
spuriously appeared to recognise all peptides; such artifactual binding to avidin has been observed in other serological studies [39].

Since the initial report of the sequence and antigenic dimorphism of TSSA by Di Noia et al [15], E. coli-produced recombinant TSSA proteins have been used as antigens with human and animal sera, as summarised in Table 3. Recognition of only TSSA-II by chronic chagasic sera from the Southern Cone region was initially interpreted as suggesting that only TcII caused chronic Chagas disease [15]. However there are many descriptions of Chagas disease and chronic cardiomyopathy in TcI endemic regions. Recognition of recombinant TSSA-I by human chagasic sera has been reported by western blot but not by ELISA [18,21]. One western blot study with recombinant TSSA-II and TSSA-I has recorded an unexpected level of TcII in northern South America and Mexico [21]. The recombinant TSSA proteins used as antigens as described encompass up to 26 amino acids flanking the polymorphic region [15,20,21], which are highly conserved between TSSA-I and TSSA-II.

The lineage-specific peptide representing the epitope common to TcII/TcV/TcVI was recognized by a large number of sera from Brazil; a proportion of these sera also bound to TSSApep-I/VI. All duplicate separate samples from the same patients gave indistinguishable results. The Brazilian sera here originated from the states of Goiás and Minas Gerais, where TcII human infections are known to be prevalent, TcV and TcVI are also present and TcI is (relatively) uncommon [11,40,41], although TcI is well represented among Brazilian sylvatic transmission cycles [41,42]. However, a substantial minority of the Brazilian serum samples (31/98 (31.6%)) did not react with TSSApep-II/VI. Thus sensitivity of the TSSApep-II/VI ELISA does not appear to be absolute for TcII/TcV/TcVI T. cruzi infections (Figure 2, Table 1). It is possible that corresponding antibodies in the TcII/TcV/TcVI seronegative patients were simply below the threshold for detection in the ELISA, although this seems unlikely because such patients remained equally seronegative against the peptides even when re-tested at the higher serum concentration of 1:100 (data not shown). Alternatively, some patients may fail to generate an immune response to the epitope or there may be as yet undiscovered TSSA diversity in some T. cruzi TcII strains.

Elsewhere in the Southern Cone countries 12 of 15 sera from Bolivia or Argentina were seropositive with TSSApep-II/VI, in accord with the known high prevalence of these lineages in those countries [23]. All sera from Bolivia, Argentina and Ecuador, and the great majority of those from Brazil, that recognised TSSApep-II/VI/V also reacted with chimera TSSApep-II-I indicating that crucial residues reside in the N-terminal part of the TSSA-II/VI epitope.

We found that few serum samples from the three countries in northern South America recognized TSSApep-II/VI/VI or TSSApep-VI. This is consistent with the literature on the geographical distribution of T. cruzi lineages based on genotyping of isolates from domestic and sylvatic transmission cycles. In fact only 4 sera from Ecuador were seropositive with TSSApep-II/VI/VI out of 66 from these northern countries. At least 3 of these 4 Ecuadorian serum samples originated from the Loja region in southern Ecuador, where TcI has been isolated [43], close to the border with Peru. Risso et al [21] reported the identification of TcII in Colombia, Venezuela, and Mexico using western blots with TSSA-II recombinant antigen. However, when the same Colombian sera samples were tested here using the lineage-specific peptides we found no TSSApep-II/VI/VI seropositive patients. Thus with the data we are unable to confirm the presence of TcII/TcV/TcVI in those Colombian patients.

Only four sera, including one from Venezuela where TcIV is known to sporadically infect humans, recognised TSSApep-IV. All four sera also recognised TSSApep-III, which shares 14 of 16 residues, presumably due to cross reaction, as we have observed with experimental murine sera (Battacharyya et al, in preparation).

Apart from one Argentine and one Brazilian serum, no clear specific reaction with TSSApep-I was observed, even with sera from known TcI endemic regions in Venezuela, Colombia and Ecuador. The few TSSApep-II/VI/VI seropositive samples from Brazil that also reacted with chimera TSSApep-II-I did not react with TSSApep-I. One possibility is that the TSSA-I protein, if expressed at all in chagasic patients, is not sufficiently immunogenic to generate an antibody response, possibly due to post-translational glycosylation of the core peptide sequence. Identification of the disaccharide Galα1,3Galβ as the immunodominant glycoepitope present in the O-linked mucins, i.e., those glycosylated on serine or threonine residues of the peptide chain, has been reported recently [44,45], and both serine and threonine are represented by one additional residue in TSSApep-I as compared

<table>
<thead>
<tr>
<th>Table 3. Reports of TSSA recombinant proteins in serological assays.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reference</td>
</tr>
<tr>
<td>[15]</td>
</tr>
<tr>
<td>[18]</td>
</tr>
<tr>
<td>[19]</td>
</tr>
<tr>
<td>[20]</td>
</tr>
<tr>
<td>[21]</td>
</tr>
<tr>
<td>[22]</td>
</tr>
<tr>
<td>[23]</td>
</tr>
</tbody>
</table>

*Chemiluminescent ELISA; *called by authors TSSA VI, but the same as that first described as TSSA-II.

doi:10.1371/journal.pntd.0002892.t003
with the TSSA-II epitope. However, equally likely, the TSSA I epitope may be conformational, with a structure that is not represented by the linear peptide. Also, alternative immunodominant epitopes elsewhere in TSSA-I may skew the humoral response away from the sequence represented by TSSApep-I.

We were interested to see whether there was a difference in the proportions of TSSApep-II/V/VI seropositive and seronegative patients presenting with clinical symptoms of chronic Chagas disease. Remarkably, there was a clear statistically significant difference: 23/60 (38%) of the Brazilian TSSApep-II/V/VI seropositives had ECG abnormalities typical of Chagas disease, whereas such abnormalities only occurred in 4/23 (17%) of the seronegatives (p<0.0001). One possible interpretation of these data is that TSSApep-II/V/VI seronegative patients may not be infected with these lineages but with some less pathogenic strains. Alternatively, such seronegative patients may be infected with TcII, TeV or TeVI but the absence of an immune response to the TSSA-II/V/VI common epitope may be an indicator of a long term better prognosis; however confirmation would require a more extensive and longitudinal study. However, the frequencies of megaeosinophagous (43% vs 48%) and megacolon (10% vs 8.7%) were not significantly different between the TSSApep-II/V/VI seropositive and seronegative groups respectively.

Trypanosoma rangeli is non-pathogenic to humans, is found sympatric with T. cruzi, particularly in northern South America, and serological cross-reaction between these species has been recognised [46]. The divergence of the TSSA epitopes in T. cruzi and the lack of response to the peptides with sera from northern South America, indicate that monospecific sera from patients infected with T. rangeli alone will not recognise these synthetic peptide epitopes.

**References**


A recent paper reported the identification of a TcI epitope for lineage-specific serology [26]. However, upon analysing the sequence diversity in the ORF coding for the parent protein across T. cruzi lineages, in contrast we found a very high degree of sequence conservation across the lineages. Thus, we were not able to confirm any TcI-specificity of that peptide epitope.

We have demonstrated that synthetic peptides are able to provide T. cruzi antigens for lineage-specific serodiagnosis in chronic Chagas’ disease. Synthetic-peptide based lineage-specific serology has also confirmed the disparate geographical distribution of TcII/TcV/TcVI but found fewer TcII infections in northern South America than reported with western blots and recombinant TSSA-II. Further comparisons of recombinant TSSA antigens and synthetic peptides are indicated. More attempts should be made to design a Tc specific peptide, and by comparative genomics to seek alternative antigens to TSSA that may be lineage-specific. However, such in silico methods will need to incorporate structural analysis, and if necessary devise linear peptides that represent conformational epitopes. In a region of Brazil endemic for TcII we find a higher rate of ECG abnormalities among patients with TSSApep-II/V/VI seropositivity than among seronegative patients. Synthetic-peptide antigens clearly have substantial and versatile potential in studying the relationship between a patient’s history of infection and clinical status, and they may provide clinical biomarkers for prognosis of Chagas disease.

**Author Contributions**

Conceived and designed the experiments: TB AKF MAM. Performed the experiments: TB AKF TTT SVL MAM. Analyzed the data: TB AKF AOL TTT SVL MAM. Contributed reagents/materials/analysis tools: AKF AOL TTT SVL MAM. Wrote the paper: TB AKF AOLO TTT SVL MAM.


28. Falconar AK (2012) Epitope reactions can be gauged by relative antibody discriminating specificity (RADS) values supported by deletion, substitution and cysteine bridge formation analyses: potential uses in pathogenesis studies. BMC Res Notes 5: 208.


