Diosque, P; Tomasini, N; Lauthier, JJ; Messenger, LA; Monje Rumi, MM; Ragone, PG; Alberti-D’Amato, AM; Pérez Brandán, C; Barnabé, C; Tibayrenc, M; +4 more... Lewis, MD; Llewellyn, MS; Miles, MA; Yeo, M; (2014) Optimized Multilocus Sequence Typing (MLST) Scheme for Trypanosoma cruzi. PLoS neglected tropical diseases, 8 (8). e3117. ISSN 1935-2727 DOI: https://doi.org/10.1371/journal.pntd.0003117

Downloaded from: http://researchonline.lshtm.ac.uk/1910909/

DOI: https://doi.org/10.1371/journal.pntd.0003117

Usage Guidelines:

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

Available under license: http://creativecommons.org/licenses/by/2.5/
Optimized Multilocus Sequence Typing (MLST) Scheme for Trypanosoma cruzi

Patricio Diosque1*, Nicolás Tomasini1, Juan José Lauthier1, Louisa Alexandra Messenger2, María Mercedes Monje Rumi1, Paula Gabriela Ragone1, Anahí Maitén Alberti-D’Amato1, Cecilia Pérez Brandán1, Christian Barnabé3, Michel Tibayrenc3, Michael David Lewis2, Martin Stephen Llewellyn2, Michael Alexander Miles2, Matthew Yeo2

1 Unidad de Epidemiología Molecular (UEM), Instituto de Patología Experimental, CONICET- Universidad Nacional de Salta, Salta, Argentina, 2 Faculty of Infectious and Tropical Diseases, Department of Pathogen Molecular Biology, London School of Hygiene and Tropical Medicine, London, United Kingdom, 3 Maladies Infectieuses et Vecteurs Ecologie, Génétique, Évolution et Contrôle, MIVEGEC (IRD 224-CNRS 5290-UM1-UM2), IRD Center, Montpellier, France

Abstract

Trypanosoma cruzi, the aetiologic agent of Chagas disease, possesses 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 characterization of different multilocus genotypes within these major lineages. Whole genome sequencing applied to large sample sizes is currently available 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 between all reference strains in the panel, with the majority supported by robust bootstrap values. Additionally, a reduced panel of just 4 gene fragments displayed high bootstrap values for DTU assignment 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.


Editor: Philippe Büsch, Institute of Tropical Medicine, Belgium

Received January 7, 2014; Accepted July 15, 2014; Published August 28, 2014

Copyright: © 2014 Diosque 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 European Union Seventh Framework Programme Grant 223034. 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.

Email: patricio.diosque@uns.edu.ar

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 distinct 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 TcIII 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 million 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
Author Summary

The single-celled parasite Trypanosoma cruzi occurs in mammals and insect vectors in the Americas. When transmitted to humans it causes Chagas disease (American trypanosomiasis) a major public health problem. T. cruzi is genetically diverse and currently split into six groups, known as Tcl to TclVI. Multilocus sequence typing (MLST) is a method used for studying the population structure and diversity of pathogens and involves sequencing DNA of several different genes and comparing the sequences between isolates. Here, we assess 13 T. cruzi genes and select the best combination for diversity studies. Outputs reveal that a combination of 7 genes can be used for both lineage assignment and high resolution studies of genetic diversity, and a reduced combination of four loci for lineage assignment. 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, genomics, ecology and epidemiology of this important human pathogen.

heterogeneity of this parasite. Subsequently, several PCR-based typing assays have been designed to differentiate the main DTUs [21,22,23,24] and more recently, combinations of PCR-RFLP schemes have been published [25,26,27]. Some approaches based on DTU characterisation by direct sequential PCR amplifications from blood and tissue samples are also promising, although various sensitivity and reliability issues need to be resolved [28,29,30]. Microsatellite typing (MLMT) has also been applied to population data for fine-scale intra DTU genetic analysis [31,32,33].

Multilocus sequence typing (MLST), originally developed for bacterial species typing, has now been applied to a wide range of prokaryotic [34,35,36,37] and increasingly eukaryotic microorganisms [38,39,40,41,42,43,44,45,46,47,48]. The technique typically involves the sequencing and concatenation of six to ten internal fragments of single copy housekeeping genes per strain [49]. Data are often hosted on interactive open access databases such as MLST.net for use in the wider research community. A major advantage of MLST analysis is that sequence data are unambiguous, minimizing interpretative errors. In this context, the MLST approach represents an excellent candidate to become the gold standard for T. cruzi genetic typing with outputs suitable for phylogenetic and epidemiological studies, particularly where large numbers of isolates from varied sources are under study.

Recently, two multilocus sequence typing (MLST) schemes have been developed in parallel for T. cruzi, each of them based on different gene targets [50,51]. Both schemes display a high discriminatory power and are able to clearly differentiate the main T. cruzi DTUs. The current work proposes to resolve the optimum combination of loci across the two schemes to produce a reproducible and robust formalised MLST scheme validated across a shared reference panel of isolates for practical use by the wider T. cruzi research community.

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-Hidroxi-3-metilgutaril-CoA reductase (HMCOAR), Piruvate dehydrogenase component E1 subunit α (PDH), Small GTP-binding protein Rab7 (GTP), Serine/tronoin-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-dihydrolate synthase (DHFR-TS), gluta-thione-dependent peroxidase II (TcGPXII), 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, HMCOAR, RHO1, GTP and STPP2. For the 13 loci under study, searches in the CL-Brener and Sylvia X10 genotypes (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 amplicon 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, JN129589-JN129590, JN129600-JN129601, JN129610-JN129617, JN129622-JN129623, JN129633-JN129634, JN129643-JN129650, JN129655-JN129656, JN129666-JN129667, JN129676-JN129683, JN129688-JN129689, JN129699-JN129700, JN129709-JN129710, JN129721-JN129722, JN129732-JN129733, JN129742-JN129749, JN129754-JN129755, JN129763-JN129765, JN129766-JN129775, JN129782-JN129783, JN129798-JN129799, JN129808-JN129815, JN129820-JN129821, KF889442-KF889646. 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: TcMARK_CONTIG_2686, TcMARK_CONTIG_670, TcMARK_CONTIG_1404, TcMARK_2068, TcMARK_3409, TcMARK_5695, TcMARK_9874, TcMARK_515, TcMARK_4904, TcMARK_5926, TcMARK_8923, TcMARK_CONTIG_1818 and TcMARK_2666.

Molecular methods

PCRs were performed in 50 μl reaction volumes containing 100 ng of DNA, 0.2 μM 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 and associated sequencing was repeated at least twice. In order to assess reproducibility, each PCR amplification was performed using the same protocols. Primers used for sequencing were identical to those used in PCR amplifications. In order to distinguish when chosen at random from a population of unrelated strains [53], was determined for each target (Table 3).

<table>
<thead>
<tr>
<th>Strain</th>
<th>DTU</th>
<th>Origin</th>
<th>Host</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. X10d1</td>
<td>Tcl</td>
<td>Belém, Brazil</td>
<td>Homo sapiens</td>
</tr>
<tr>
<td>2. Cutia cl1</td>
<td>Tcl</td>
<td>Espirito Santo, Brazil</td>
<td>Dasyprocta aguti</td>
</tr>
<tr>
<td>3. Sp104 cl1</td>
<td>Tcl</td>
<td>Region IV, Chile</td>
<td>Triatoma spinalei</td>
</tr>
<tr>
<td>4. P209 cl93</td>
<td>Tcl</td>
<td>Sucre, Bolivia</td>
<td>Homo sapiens</td>
</tr>
<tr>
<td>5. OPS21 cl11</td>
<td>Tcl</td>
<td>Cojedes, Venezuela</td>
<td>Homo sapiens</td>
</tr>
<tr>
<td>6. 92101601 cl1</td>
<td>Tcl</td>
<td>Georgia, USA</td>
<td>Didelphis marsupialis</td>
</tr>
<tr>
<td>7. TU18 cl93</td>
<td>Tcl</td>
<td>Potosí, Bolivia</td>
<td>Triatoma infestans</td>
</tr>
<tr>
<td>8. CBB cl3</td>
<td>Tcl</td>
<td>Region IV, Chile</td>
<td>Homo sapiens</td>
</tr>
<tr>
<td>9. Mas cl1</td>
<td>Tcl</td>
<td>Federal District, Brazil</td>
<td>Homo sapiens</td>
</tr>
<tr>
<td>10. IVV cl4</td>
<td>Tcl</td>
<td>Region IV, Chile</td>
<td>Homo sapiens</td>
</tr>
<tr>
<td>11. Esm cl3</td>
<td>Tcl</td>
<td>São Felipe, Brazil</td>
<td>Homo sapiens</td>
</tr>
<tr>
<td>12. M5631 cl5</td>
<td>Tcl</td>
<td>Selva Terra, Brazil</td>
<td>Dasyus novemcinctus</td>
</tr>
<tr>
<td>13. M6241 cl6</td>
<td>Tcl</td>
<td>Belem, Brazil</td>
<td>Homo sapiens</td>
</tr>
<tr>
<td>14. CM17</td>
<td>Tcl</td>
<td>Meta, Colombia</td>
<td>Dasyur sp.</td>
</tr>
<tr>
<td>15. X109/2</td>
<td>Tcl</td>
<td>Makthlawaiya, Paraguay</td>
<td>Canis familiaris</td>
</tr>
<tr>
<td>16. 92122102R</td>
<td>TclV</td>
<td>Georgia, USA</td>
<td>Procyn lotor</td>
</tr>
<tr>
<td>17. Can3 cl1</td>
<td>TclV</td>
<td>Belém, Brazil</td>
<td>Homo sapiens</td>
</tr>
<tr>
<td>18. Dog Theis</td>
<td>TclV</td>
<td>USA</td>
<td>Canis familiaris</td>
</tr>
<tr>
<td>19. Mn cl2</td>
<td>TclV</td>
<td>Region IV, Chile</td>
<td>Homo Sapiens</td>
</tr>
<tr>
<td>20. Bug 2148 cl1</td>
<td>TclV</td>
<td>Rio Grande do sul, Brazil</td>
<td>Triatoma infestans</td>
</tr>
<tr>
<td>21. SO3 cl5</td>
<td>TclV</td>
<td>Potosi, Bolivia</td>
<td>Triatoma infestans</td>
</tr>
<tr>
<td>22. SC43 cl1</td>
<td>TclV</td>
<td>Santa-Cruz, Bolivia</td>
<td>Triatoma infestans</td>
</tr>
<tr>
<td>23. CL Brener</td>
<td>TclV</td>
<td>Rio Grande do Sul, Brazil</td>
<td>Triatoma infestans</td>
</tr>
<tr>
<td>24. P63 cl1</td>
<td>TclV</td>
<td>Makthlawaiya, Paraguay</td>
<td>Triatoma infestans</td>
</tr>
<tr>
<td>25. Tula cl2</td>
<td>TclV</td>
<td>Talahuen, Chile</td>
<td>Homo sapiens</td>
</tr>
</tbody>
</table>

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

Sequence data were concatenated and Neighbour Joining method 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
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). STPP2 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 Mncl2 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 states) 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 (3,177 combinations) by implementing the MLSTest scheme optimisation algorithm. One combination of 7 loci was the best according to the proposed criteria. This combination consisted of Rb19, TcMPX, HMCOAR, RHO1, GPI, SODB and LAP discriminating all 25 strains as DSTs, and importantly categorising all separate DTUs as a monophyletic group. DTUs were also well-supported by associated bootstraps values (TcI,100; TcII, 100; TcIII, 99.8; TcIV, 88.2; TcV, 88.7; TcVI, 99.6) as illustrated in Figure 2. Combinations with higher number of loci (from 8 to 12) did not significantly increased bootstrap values of TcIV and TcV.

We assessed whether the outlier for TcV (Mncl2) and the low bootstrap observed for TcVI (applied to all 13 fragments) was due to incongruence among fragments. The thirteen fragment dataset was significantly incongruent (BIONJ-ILD p-value<0.001) for at least one partition which was corroborated using NJ-ILD with a permutation test and 500 replicates. Significant incongruence (p-value<0.05 after Bonferroni correction) was detected in the TcV and TcVI nodes. Incongruence was likely due to strains within DTUs TcV and TcVI demonstrating apparent loss of heterozygosity (LOH) in the hybrid DTU TcV, retaining loss of heterozygosity (LOH) in the hybrid DTU TcV, retaining loss of heterozygosity (LOH) in the hybrid DTU TcV, retaining loss of heterozygosity (LOH) in the hybrid DTU TcV, retaining loss of heterozygosity (LOH) in the hybrid DTU TcV, retaining.

### 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>RHO1*</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>0.608</td>
<td>0.985</td>
</tr>
<tr>
<td>GPY</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].
†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

Revised 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, RHO1 and GPI (four loci) produced the highest bootstrap values for DTU assignment across the DTUs, TcI (99.9), TcII (100), TcIII (99.5), TcIV (96.7), 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, RHO1 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-TcIII-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 Mncl2 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, and 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
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 *T. 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 DSTs 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

---

**Figure 1.** Neighbor Joining tree based on the concatenation of 13 MLST fragments. Different DTUs are represented by vertical bars. Branch values represent bootstrap values (1000 replications), different bootstrap values indicate the method of handling heterozygous sites: SNP duplication method (first value) and average states (second values). Branch supports highlighted in blue shows branches where support for SNP duplication method was higher than the average states method. The outlier TcV is highlighted in red. Scale bar at the bottom left represents uncorrected p-distances. doi:10.1371/journal.pntd.0003117.g001
the physical differences between TcV and TcVI. As the phylogenetic incongruence is genetic recombination, one interpretation of this is that TcI has a closer affinity to TcIII than to TcIV. TcII representatives revealed clustering of DTUs TcI, TcIII and TcIV, indicating that TcI has 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 TrMPX 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 TrMPX sequencing due to apparent LOH. Despite this LOH the TrMPX locus was included in the 4 target scheme to increase bootstrap support in differentiating between TcV and TcVI.

Although MLST has been successfully applied to other diploid organisms including Candida albicans, the potential for heterozygous alleles complicates typing schemes. In the present work, two methods to handle heterozygous sites, SNPs duplication and average states algorithms, produced broadly similar results with SNP duplication producing marginally higher bootstraps due to the physical duplication of informative sites. Here we decided to implement the average states methodology to derive genetic distances and phylogenies. Both approaches can be found in the software MLSTest [52] producing results that enable resolution at the DTU level and an associated DP of 1 for the panel tested. A significant advantage of MLST based analysis over sequential PCR based gels is that once generated, sequences can be applied to a range of complementary downstream analyses. For example, the resolution of haplotypes for recombination analysis and investigation of more detailed evolutionary associations can be applied to population sized studies. At present, whole genome sequencing applied to large numbers of isolates is not feasible and microsatellite analysis is often difficult to reproduce precisely across laboratories, unlike MLST which has proven reproducibility both within and between laboratories [59]. However, microsatellites could be more convenient for population genetics studies at a microevolutionary level, due to their high resolution power. A further consideration in the analysis of diploid sequences is the likelihood of recombination since full genome sequencing applied to large numbers of isolates is not feasible. Microsatellite analysis is often difficult to reproduce precisely across laboratories, unlike MLST which has proven reproducibility both within and between laboratories [59]. However, microsatellites could be more convenient for population genetics studies at a microevolutionary level, due to their high resolution power.

Figure 2. Neighbor Joining tree based on the concatenation of 7 selected MLST fragments: Rb19, TcMPX, HMCOAR, RH01, GPI, SOD2 and LAP. Different DTUs are represented by vertical bars. Branch values represent bootstrap values (1000 replications). Heterozygous sites were considered as average states (see methods). Scale bar at the bottom left represents uncorrected p-distances.

Figure 3. Neighbor Joining tree based on the concatenation of 4 selected MLST fragments (TCMPX, HMCOAR, RH01, GPI) for DTU assignment. Different DTUs are represented by vertical bars. Branch values represent bootstrap values (1000 replications). Heterozygous sites were handled using the average states method. Scale bar at the bottom left represents uncorrected p-distances.
References


