Phylogeography of Toxoplasma gondii Points to a South American Origin

Emilie Bertranpetit, Thibaut Jombart, Emmanuel Paradis, Hilda Pena, Jitender Dubey, Chunlei Su, Aurélien Mercier, Sébastien Devillard, Daniel Ajzenberg

aINSERM UMR_S 1094, Neuroépidémiologie Tropicale, Laboratoire de Parasitologie-Mycologie, Faculté de Médecine, Université de Limoges, Limoges, 87025, France
bMRC Centre for Outbreak Analysis and Modelling, Department of Infectious Disease Epidemiology, School of Public Health, Imperial College London, United Kingdom
cInstitut des Sciences de l’Évolution, Université Montpellier/CNRS/IRD/EPHE, Place Eugène Bataillon – CC 065, 34095 Montpellier cédex 05, France
dDepartamento de Medicina Veterinária Preventiva e Saúde Animal, Faculdade de Medicina Veterinária e Zootecnia, Universidade de São Paulo, São Paulo, Brazil
eUnited States Department of Agriculture, Agricultural Research Service, Beltsville Agricultural Research Center, Animal Parasitic Diseases Laboratory, Building 1001, Beltsville, Maryland, 20705-2350, USA
fDepartment of Microbiology, University of Tennessee, Knoxville, Tennessee, 37996-0845, USA
gUniversité de Lyon, F-69000, Lyon ; Université Lyon 1 ; CNRS, UMR 5558, Laboratoire de Biométrie et Biologie Evolutive, F-69622, Villeurbanne, France
hCentre National de Référence (CNR) Toxoplasmose / Toxoplasma Biological Resource Center (BRC), Centre Hospitalier-Universitaire Dupuytren, Limoges, 87042, France and INSERM UMR_S 1094, Neuroépidémiologie Tropicale, Laboratoire de Parasitologie-Mycologie, Faculté de Médecine, Université de Limoges, Limoges, 87025, France

Correspondence and requests for materials should be addressed to S.D (email: sebastien.devillard@univ-lyon1.fr Tel: +33472448111) and D.A. (email: ajz@unilim.fr Tel: +33555056160)

Authors contributed equally
Abstract

Toxoplasma gondii, a protozoan found ubiquitously in mammals and birds, is the etiologic agent of toxoplasmosis, a disease causing substantial Public Health burden worldwide, including about 200,000 new cases of congenital toxoplasmosis each year. Clinical severity has been shown to vary across geographical regions, with South America exhibiting the highest burden. Unfortunately, the drivers of these heterogeneities are still poorly understood, and the geographical origin and historical spread of the pathogen worldwide are currently uncertain. A worldwide sample of 168 T. gondii isolates gathered in 13 populations was sequenced for five fragments of genes (140 single nucleotide polymorphisms from 3,153 bp per isolate). Phylogeny based on Maximum likelihood methods with estimation of the time to the most recent common ancestor (TMRCA) and geostatistical analyses were performed for inferring the putative origin of T. gondii. We show that extant strains of the pathogen likely evolved from a South American ancestor, around 1.5 million years ago, and reconstruct the subsequent spread of the pathogen worldwide. This emergence is much more recent than the appearance of ancestral T. gondii, believed to have taken place about 11 My ago, and follows the arrival of felids in this part of the world. We posit that an ancestral lineage of T. gondii likely arrived in South America with felids and that the evolution of oral infectivity through carnivorism and the radiation of felids in this region enabled a new strain to outcompete the ancestral lineage and undergo a pandemic radiation.

Keywords: Toxoplasma gondii, phylogeography, Maximum likelihood phylogeny, time to the most recent common ancestor (TMRCA), genetic diversity
1. Introduction

Toxoplasmosis imposes a substantial disease burden across the world. Serological studies demonstrate its presence in virtually every country, with seroprevalence exceeding 60% in some parts of South America, Africa, and South-East Asia (Pappas et al., 2009). While asymptomatic in most patients, toxoplasmosis is a major cause of uveitis in immunocompetent patients and a potentially life-threatening illness in immunocompromised patients and fetuses (Montoya and Liesenfeld, 2004). Congenital toxoplasmosis alone has a yearly global incidence of about 200,000 cases, causing a burden exceeding 1 million disability-adjusted life year (DALYs) (Torgerson and Mastroiacovo, 2013). The overall disease burden attributable to all forms of toxoplasmosis is certainly far greater, and highest in South America where ocular toxoplasmosis is unusually frequent and severe (Glasner et al., 1992; de-la-Torre et al., 2008; Gilbert et al., 2008; Torgerson and Mastroiacovo, 2013).

The extent to which host parasite genetics, host immune status, and exposure rate contribute to the increased severity of toxoplasmosis in South America is unclear, but differences in the genetic makeup of *T. gondii* strains are likely to play a major role (Khan et al., 2006; Gilbert et al., 2008). In North America, Europe, Africa, and Asia, the population structure of *T. gondii* is dominated by a few prevalent clonal strains, whereas much greater genetic diversity is seen in tropical South America where the populations lack sign of recent genetic bottleneck and clonal structure seen in the other parts of the world (Shwab et al., 2014; Lorenzi et al., 2016).
Unfortunately, the drivers of the pathogen’s genetic diversity are still poorly understood, and the origin of extant lineages of *T. gondii* remains controversial. Recent work suggested a potential South American origin, while another study advocated that co-migration with felids led to the divergence of South American strains from pre-existing North American ones ([Lehmann et al., 2006; Khan et al., 2007](#)). The estimation of the time to the most recent common ancestor (TMRCA) of extant lineages is also disputed, with estimates ranging from 150,000 to $10^7$ years ([Morrison, 2005; Khan et al., 2007](#)).

In the present study, using a large collection of genetic sequences of *T. gondii* sampled worldwide, we reconstructed the phylogeography of *T. gondii* as a basis to address the controversial questions regarding the evolution of this parasite and its geographical origin.

### 2. Materials and Methods

#### 2.1. Collection of *T. gondii* strains and selection of markers

A total of 168 *T. gondii* strains from 13 populations collected worldwide in North America, South America, the Caribbean, Europe, Asia, and Africa, were used in this study (Supplementary information and S1 Table). Our collection was specifically designed to ensure extensive geographic coverage of strains clustered in true populations, and includes a large number of strains from Africa and China, which were so far underrepresented or absent in the previous studies ([Lehmann et al., 2006; Khan et al., 2007](#)). A preliminary genetic analysis with 15 microsatellite markers (Ajzenberg et al., 2010) was performed to exclude clones of strains sampled in the same area. Each isolate was sequenced in both directions for five markers (*GRA6, GRA7, SAG3, UPRT1*)
and UPRT7) that had the highest polymorphic rates after a preliminary analysis of genetic polymorphism of 30 fragments of genes retrieved from GenBank and ToxoDB (Supplementary information and S2 Table).

2.2. Phylogenetic analysis.

*Hammondia hammondi* is the most closely parasite related to *T. gondii* and was used as an outgroup in phylogenetic analyses. Sequences of the *H. hammondi* strain H.H.34 corresponding to *T. gondii* GRA6, GRA7, SAG3, UPRT1 and UPRT7 sequences were retrieved from GenBank and ToxoDB, and aligned with MUSCLE (Edgar, 2004). We used ape and pegas R packages to extract haplotypes and build phylogenetic trees with three distance-based methods: NJ, BIONJ, and FastME which were used as starting trees for the ML analyses. Maximum likelihood phylogenetic analyses were performed with the R package phangorn using four partitions of the sequence data by crossing two criteria: exons vs. introns on one side, and GRA6, GRA7 and SAG3 vs. UPRT1 and UPRT7 on the other (Supplementary information). A GTR + I + I model was used with parameters that could vary among data partitions. The different model fits performed with phangorn were compared with AIC. The three trees obtained with the distance-based methods were all tested as initial trees.

2.3. Geostatistical analyses. We used the same geostatistics approach for inferring the putative origin of *T. gondii* as previously used for uncovering the origins of *P*. 
*falciparum* (Tanabe et al., 2010, 2013a, 2013b; Mita and Jombart, 2015). We implemented this approach in the R package *geoGraph* ([http://thibautjombart/geograph](http://thibautjombart/geograph)), in which we provided extensive documentation replicating the analyses described below using publicly available data (Cann et al., 2002). The method implemented in *geoGraph* relies on the idea that migration events result in successive bottlenecks which reduce the genetic diversity within populations as they are located further away from the origin (Tanabe et al., 2010). Accordingly, we expect to observe a negative correlation between within-population diversity and the distance from the origin. While in practice the true origin is often unknown, one can infer the most plausible origin by assessing this relationship for a number of candidate origins, and retaining the origin yielding the strongest negative correlation. This method requires two types of distances, genetic and spatial, to be computed. Here, the genetic diversity was mostly structured by varying frequencies of a small number of haplotypes within populations (S3 Table and Fig 1). Therefore, we used haplotype richness (i.e. number of distinct haplotypes) as a measure of diversity within populations. Spatial distances through landmasses were computed using *geoGraph*. The package models movements on the surface of the Earth using a spherical, pseudo-regular grid with approximately 40,000 nodes. Each node possesses an ‘habitat’ attribute, here used to distinguish landmasses from seas. Shortest path between locations were computed using the dijkstra algorithm (Jungnickel, 2013) implemented in the R package RBGL (Edmonds et al., 2006; Carey et al., 2011).
define candidate origins, 1,800 combinations of regularly spaced longitudes and
latitudes were used to cover the globe, which resulted in 433 non-redundant locations
on landmasses on the grid used by geoGraph. For each location, the shortest path
through each sampled population was identified, and the corresponding distance
computed in kilometers. These distances were then used to assess patterns of
decrease of genetic diversity from the putative origin using simple linear regression. The
most likely origin was inferred as the location which yielded the most negative
correlation between geographic distances and haplotype diversity within populations.

2.4. Time to the most recent common ancestor (TMRCA). In order to estimate the
time to the most recent common ancestor (TMRCA) of T. gondii, we used two different
approaches: a simple molecular dating method based on the divergence with H. hammondi which is estimated to be around 11 My, and a coalescent approach using the
expectation of TMRCA which is equal to twice the effective population size (Ne). Both
approaches need an estimate of the mutation rate (μ), and the second one also needs
an estimate of the population parameter θ (= 2 μ Ne). We estimated μ for the non-
coding introns of UPRT1 and UPRT7. We did two neutrality tests: D’s Tajima and the
R2 test. We calculated a standard-error of this estimate using the variance of ratio of
two random variables, here the number of mutations per site and twice the time of
divergence. The former was estimated with a Tamura-Nei distance with its associated
variance, and the latter was 11 My with an arbitrary sd = 1 My. The population parameter θ was estimated in two ways: with a Markov chain Monte Carlo (MCMC) approach as implemented in the R package coalescentMCMC, and with the nucleotide diversity (π) calculated with pegas. Both ways calculate the standard-error of the estimate of θ. In the end, three estimates of TMRCA were obtained with their respective 95% confidence interval (CI).

3. Results

3.1. Genetic diversity.

Sequences of the five markers represented a total of 3,153 bp per isolate, including 140 variable sites. Without taking into account sites with gaps, 26, 30, 27, 32, and 25 SNPs were identified in the GRA6 (607 bp), GRA7 (677 bp), SAG3 (638 bp), UPRT1 (574 bp), and UPRT7 (657 bp) genes, respectively (S4 Table and S1 Fig). Strains from the Caribbean, Guiana shield, Northeast Brazil, and Southeast Brazil showed more sequence polymorphism with 65, 90, 59, and 65 SNPs, respectively (S3 Table). Polymorphism was lower in the African and European populations with a number of SNPs ranging from one to 39. Polymorphism was intermediate in the Asian and North American populations with 41, 60, 47, and 51 SNPs in Turkey, China, Minnesota, and Pennsylvania, respectively.
Data concatenation revealed 60 haplotypes. The number of haplotypes ranged from four in Europe to 32 in South America. Of the 32 genotypes in the 44 strains from South America, 29 were endemic in South America whereas only three were common in other populations (one in the Caribbean, one in Africa, and one both in Asia and Africa). In contrast, of the four genotypes in the 34 strains from Europe, three were common in other populations (one in Asia and Africa, one in North America, Asia and Africa, and one in the Caribbean, North America, Asia and Africa) and the unique genotype differed by only one SNP from the one common to North America, Asia, and Africa (S5 Table).

The higher values of haplotype diversity, estimated from concatenation of the five markers by the number of haplotypes divided by the number of isolates, were observed in South America and the Caribbean, whereas the lower values were observed in Europe, Asia, and Africa (S3 Table). Haplotype diversity was intermediate in North America. Overall the highest genetic diversity was found in South-America.

3.2. Phylogeny of *T. gondii* strains

Preliminary analyses of genetic diversity (Supplementary information, S6 Table and S2 and S3 Figs.) revealed different mutation patterns and rates of evolution between coding and non-coding segments, and for the two *UPRT* genes compared to the others. Accordingly, we defined four partitions of the sequence data crossing these two categories, and reconstructed separate phylogenies by maximum likelihood (ML) to investigate potential phylogenetic incongruence (Som, 2015). Statistical tests and
examination of model selection criterion (AIC) revealed the existence of distinct
topologies (Fig. 1), suggesting that these sequence partitions have undergone different
evolutionary histories and selective pressures. Interestingly, only South American
strains were consistently placed at a basal position (close to the root) in all topologies.
To investigate this pattern further and identify the common evolutionary history of these
genes, a consensus topology was inferred from the four ML-partitioned topologies (Fig.
2). This new tree supported the more ancestral status of South American isolates, with
35 out of 44 samples located at the root of the tree. However, as expected in the
presence of conflicting phylogenetic signal, this tree was only partially resolved, and
strains from other locations (China: 12 samples; Africa: 7 samples) also belonged to the
large basal multifurcation.

3.3. On the geographic origin of *T. gondii*

As a complementary analysis, we used a geostatistical approach previously employed
for identifying the origin of *Plasmodium falciparum*, the main etiologic agent of malaria
(Tanabe et al., 2010, 2013b; Mita and Jombart, 2015). This method identifies likely
geographic origins as the locations from which patterns of decrease in genetic diversity,
expected to be observed due to repeated migration and founder effects, are most
consistent (Tanabe et al., 2010). Because of the low level of polymorphism observed in
the sequenced genes and the highly clonal nature of *T. gondii*, haplotype richness was
used as a measure of genetic diversity within populations (S3 Table). Testing a large
number of hypothetical origins across the world, this approach identified South America,
and more specifically Colombia as the most likely origin ($r=-0.81$, $p=0.9\times10^{-4}$, Fig. 3).
While substantial uncertainty remains about the exact location, this analysis brings
strong support to a South American origin for *T. gondii* suggested by the phylogenetic approach. Our results further suggest that *T. gondii* initially spread through the Americas and then colonized Asia and Europe via the Bering Strait, before entering Africa through two different migration routes (Fig. 3).

### 3.4. Time to the most recent common ancestor (TMRCA) of *T. gondii*.

To understand the processes which may have led to a pandemic radiation of *T. gondii* from South America, the emergence of this ancestral, highly successful lineage has to be dated. To this end, we derived estimates of time to the most recent common ancestor (MRCA) of the extant strains using standard molecular approaches. While confidence intervals indicated substantial uncertainty, overall results suggest that the MRCA of *T. gondii* appeared around 1.5 My ago (Table 1). This emergence is much more recent than the existence of *T. gondii* itself, estimated to have diverged from its closest ancestor *Hammondia hammondi* some 11 My ago (Morrison et al., 2004).

### 4. Discussion

A major event occurred in the evolutionary history of *T. gondii* which led to a selective sweep about 1.5 My ago. We hypothesized that an ancestral form of *T. gondii* was introduced in South America through the migration of Felidae after the emergence of the Isthmus of Panama about 2 to 3 My ago, at the end of Pliocene (O’Brien et al., 2008). It is believed that Felidae species quickly expanded after their arrival and diversified into the “ocelot” lineage in South America. Interestingly the Muridae, potential intermediate hosts for *T. gondii*, also showed extensive diversification in South America with the
appearance of several genera (Webb, 2006) around the same time. As previously
suggested (Webb, 2006), we posit that this expansion, diversification and mixing of host
populations certainly resulted in similar processes in their parasites and favoured the
accumulation of genetic diversity in *T. gondii*, which eventually led to a selective sweep
by a highly successful mutant lineage.

The selective pressures underlying this selective sweep can be debated. Previous work
attributed this radiation to the emergence of transmission through carnivorism (i.e., oral
infectivity of tissue cysts) between intermediate hosts in clonal strains 10,000 years ago
(Su et al., 2003). However, oral infectivity was shown to be also a trait of many South
American strains (Carme et al., 2002; Khan et al., 2007). Because South American strains
were the first to diverge from the MRCA, it is likely that transmission by carnivorism
evolved earlier than the apparition of clonal lineages. This trait conferred a better
transmission of the current form of *T. gondii* which likely outcompeted the ancestral form
arriving in South America with Felidae. The transmission of *T. gondii* between its
different hosts would allow some genotypes to migrate to North America, then to go
through the Bering Strait to colonize Asia, Europe and Africa. The current population
structure of *T. gondii* with a predominance of a few successful clonal strains in Africa,
Asia and Europe, is likely to be the consequence of the recent expansion of the
domestic cat, an Old World species until the sixteenth century, that tremendously
amplified a specific subset of pre-adapted genotypes (Müller and Howard, 2016).

In addition to being the likely origin of modern *T. gondii* strains, South America also
suffers from the highest burden of toxoplasmosis. Prevalence, incidence, and severity of
acquired and congenital ocular toxoplasmosis (OT) in some areas of Brazil, Colombia, and Argentina are considerably higher than anywhere else, which makes OT a genuine public health issue in South America (Glasner et al., 1992; de-la-Torre et al., 2008; Gilbert et al., 2008; Rudzinski et al., 2016). Because South America is also the hotspot of *T. gondii* genetic diversity, it has been hypothesized that severe forms of toxoplasmosis may be the consequence of poor adaptation of the human host to the unusual diversity of strains in this part of the world, resulting in impaired immune response and, thus, a more aggressive disease (Khan et al., 2006; Gilbert et al., 2008; Demar et al., 2012; de-la-Torre et al., 2013; Rudzinski et al., 2016). The societal and economic costs of care for symptomatic cases of congenital toxoplasmosis can be considerable but the cost-effectiveness of national routine prenatal screening and treatment program are still debated (Wallon et al., 1999; Jones et al., 2014). There is a need for randomized placebo-controlled trials to help determine the effectiveness of these interventions.

5. Conclusion

Our reconstruction of *T. gondii*’s phylogeography provides a new framework for understanding patterns of genetic diversity in sampled populations of the parasite, and for predicting diversity in unsampled locations. Because genetic diversity seems to impact directly the severity of the disease, our results can be used as a basis for explaining geographic heterogeneities in disease burden, and identifying priority targets for potential future interventions.
ACKNOWLEDGEMENTS

We are thankful to github (http://github.com/) and travis (http://travis-ci.org/) for providing great resources for the development of the R package geoGraph. We are indebted to Endrias Zewdu Gebremedhin, Hüseyin Can, Mert Döşkaya, Yüksel Gürüz, Susana Sousa, Lénaïg Halos, Min Li, François Peyron, and the members of the French National Reference Center for toxoplasmosis (Nicole Desbois, Saadia Azi, Hélène Yera, Magalie Demar, Bernard Carme, Stéphane Simon, Denis Blanchet, Rachida Boukhari, Isabelle Villena, Dominique Aubert) for their contribution to the collection of 168 strains that were included in the analysis.

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

The sequences reported in this study have been deposited in the GenBank database (accession n° KU598987-KU599154 for GRA6, accession n°KU599155-KU599322 for GRA7, accession n°KU599323-KU599490 for SAG3, accession n°KU599491-KU599658 for UPRT1 and accession n°KU599659-KU599826 for UPRT7).
References


Table 1. Estimates of the time to the most recent common ancestor (TMRCA) of *Toxoplasma gondii* with three different methods on the introns of *UPRT* genes.

<table>
<thead>
<tr>
<th>Method</th>
<th>TMRCA (Ma)</th>
<th>[95% CI]</th>
</tr>
</thead>
<tbody>
<tr>
<td>Molecular dating</td>
<td>1.59</td>
<td>[0.00–3.46]</td>
</tr>
<tr>
<td>Coalescent</td>
<td>1.26</td>
<td>[0.94–1.57]</td>
</tr>
<tr>
<td>Nuc. div. ((\pi))</td>
<td>1.20</td>
<td>[0.00–2.44]</td>
</tr>
</tbody>
</table>

Nuc. div.: nucleotide diversity.