Admixture and recombination among Toxoplasma gondii lineages explain global genome diversity

The MIT Faculty has made this article openly available. Please share how this access benefits you. Your story matters.

<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>As Published</td>
<td><a href="http://dx.doi.org/10.1073/pnas.1117047109">http://dx.doi.org/10.1073/pnas.1117047109</a></td>
</tr>
<tr>
<td>Publisher</td>
<td>National Academy of Sciences (U.S.)</td>
</tr>
<tr>
<td>Version</td>
<td>Final published version</td>
</tr>
<tr>
<td>Accessed</td>
<td>Sat Jun 17 10:20:30 EDT 2017</td>
</tr>
<tr>
<td>Citable Link</td>
<td><a href="http://hdl.handle.net/1721.1/77588">http://hdl.handle.net/1721.1/77588</a></td>
</tr>
<tr>
<td>Terms of Use</td>
<td>Article is made available in accordance with the publisher's policy and may be subject to US copyright law. Please refer to the publisher's site for terms of use.</td>
</tr>
<tr>
<td>Detailed Terms</td>
<td></td>
</tr>
</tbody>
</table>
Admixture and recombination among *Toxoplasma gondii* lineages explain global genome diversity

Samuel Minostra,b,1, Mariane B. Melo,a,c, Fugen Lid, Diana Lu,e, Wendy Niedelmana, Stuart S. Levinea, and Jeroen P. J. Saeijb,2

* Biology Department, Massachusetts Institute of Technology, Cambridge, MA 02139; and b Department of Microbiology, University of Pennsylvania Perelman School of Medicine, Philadelphia, PA 19104

Edited by Louis H. Miller, National Institutes of Health, Rockville, MD, and approved July 10, 2012 (received for review October 19, 2011)

*Toxoplasma gondii* is a highly successful protozoan parasite that infects all warm-blooded animals and causes severe disease in immunocompromised and immune-naive humans. It has an unusual global population structure: In North America and Europe, isolated strains fall predominantly into four largely clonal lineages, but in South America there is great genetic diversity and the North American clonal lineages are rarely found. Genetic variation between *Toxoplasma* strains determines differences in virulence, modulation of host-signaling pathways, growth, dissemination, and disease severity in mice and likely in humans. Most studies on *Toxoplasma* genetic variation have focused on either a few loci in many strains or low-resolution genome analysis of three clonal lineages. We use whole-genome sequencing to identify a large number of SNPs between 10 *Toxoplasma* strains from Europe and North and South America. These were used to identify haplotype blocks (genomic regions) shared between strains and construct a *Toxoplasma* haplotype map. Additional SNP analysis of RNA-sequencing data of 26 *Toxoplasma* strains, representing global diversity, allowed us to construct a comprehensive genealogy for *Toxoplasma gondii* that incorporates sexual recombination. These data show that most current isolates are recent recombinants and cannot be easily grouped into a limited number of haplogroups. A complex picture emerges in which some genomic regions have not been recently exchanged between any strains, and others recently spread from one strain to many others.

evolution | haplotype | pathogen | selection

The protozoan, *Toxoplasma gondii*, a highly prevalent parasite of warm-blooded animals including humans, is an excellent model to study sexual recombination and its effects on the population structure of eukaryotic unicellular pathogens. It has two distinct reproductive mechanisms: (i) asexual reproduction in intermediate hosts, in which tissue cysts containing haploid parasites infect other hosts through carnivores and omnivores; and (ii) sexual recombination in felines, the definitive host. The latter is more productive—an infected feline sheds millions of extremely stable, infectious oocysts (1). Because a single parasite can produce both micro- and macrogametes and self-mate, novel recombined genotypes only form in the rare event that a feline is recombined genotypes from 1 to 3%, and variation within a lineage is very low (<0.01%), suggesting recent expansion (5). A fourth clonal lineage prevalent in NA wild animals, haplogroup (HG) 12, was recently described (6). Strains with atypical or novel allele combinations have also been isolated from patients with unusual clinical presentations (7) or on other continents, particularly South America (SA) (8). The causes of these geographically confined population structures are unknown. Because some (usually SA) strains can cause severe disease even in immunocompetent humans (7), identifying the factors affecting the movement and spread of certain strains is of the utmost importance.

Evidence supports a model of frequent sexual recombination followed by strong selective sweeps. Genome-wide SNP comparison of clonal-lineage strains determined that the ancestor of type II (II*) crossed with ancestral strains α and β ~10,000 y ago to generate types I and III, respectively (9–11). Types I and III are frequently isolated in E and NA, suggesting that they have spread rapidly since their recent creation. It is unclear why types I and III and no other progeny were so prolific, but a likely explanation is that recombination brought together a specific combination of alleles that provided them with a selective advantage (12). The clonal lineages share a common chromosome Ia (chrIa) (9, 13) and a small segment of the distal part of chrXI (9), suggesting these regions might be especially advantageous. Although the genetic determinants remain elusive, we know the clonal lineages are extremely successful, albeit limited to E and NA. Surprisingly, many SA isolates carry a chrIa nearly identical to that of the clonal lineages (8, 11).

More recently, SNP analysis of eight introns in five genes from many—including atypical and SA—isolates sorted all strains into 14 major HGs, with the clonal lineages designated haplotypes I, II, and III and atypical strains distributed among HGs 4 through 14 (11). Further analyses indicated that all 14 HGs may have derived from admixture of 6 ancestral strains (11). These 14 HGs have distinct geographical distributions: I–III, 7, 11, and 12 are mainly confined to E and NA; 4, 5, 8, 9, and 10 to SA; and 13 to Asia (14). HG6 is found worldwide (11).

A drawback of the previous analyses that led to the current evolutionary model is its reliance on sequence information of only five loci. Moreover, the sequences were concatenated and analyzed as if they share a single evolutionary history. However, *Toxoplasma* undergoes sexual recombination, so different loci may have completely different evolutionary histories (15). We argue that evolutionary analysis of *Toxoplasma* (or any sexual organism) can only be achieved by sequencing a large number of loci, determining which loci have a shared recombinational history, then estimating the evolutionary history of each of those haplotype blocks independently. Analysis of haplotype blocks is greatly aided by the fact that *Toxoplasma* only replicates as a haploid.


The authors declare no conflict of interest.

This article is a PNAS direct submission.

Data deposition: The sequence reported in this paper has been deposited in Short Read Archive, http://www.ncbi.nlm.nih.gov/sra (accession nos. SRA047110 and SRA050293).

1S.M. and M.B.M. contributed equally to this work.

2To whom correspondence should be addressed. E-mail: jsaeij@mit.edu.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1117047109/-/DCSupplemental.
We used whole-genome SNP analysis of 26 *Toxoplasma* strains, representing global diversity, to construct a haplotype map and determine phylogeny for all haplotype blocks independently. Our results show that the observed extent of *Toxoplasma* genetic diversity cannot be explained by the mating of four to six ancestral strains, nor do most strains fit into the 14 proposed haplogroups. Instead, most strains appear to have formed through recent recombination events. Some of these recombinations likely led to particularly fit genotypes that geographically spread or swept. Particularly, three haploblocks, on chrIa, XI, and XII, have recent common ancestry (RCA) across multiple diverse lineages, perhaps suggesting that these regions may contain fitness loci. Our approach improves upon current methods because it analyzes recombination-defined haploblocks, building a more precise and comprehensive genealogy of *Toxoplasma*.

**Results**

**Pairwise Strain Comparisons of SNP Frequencies Identify Haplotype Blocks.** Recombination between any two strains results in progeny whose genomes are a mosaic of the parental genomes. We obtained the genome sequences of seven *Toxoplasma* strains representing HGs 4 through 11, except 7, and public genome data for the clonal lineages [Toxodb.org (16)]. To identify genomic regions with RCA, we made genomewide relative SNP rate plots for the 45 pairwise comparisons between the 10 strains. We will focus our initial analysis on the comparisons between strains that have previously been suggested to be related to each other based on the haplogroup mixture model (8) (Fig. 1). These SNP rate plots show distinct SNP patterning on each chromosome that demarcates regions (Dataset S1 A–D) shared between certain strains (Fig. S1 A–D). For example, it has been suggested that a cross between type II* and β (HG9-like) created the type III lineage (Fig. 1C) (9, 11). If this were the case, one would expect large regions in the type III genome to be very similar to either type II or HG9. From the pairwise SNP rates between ME49 (type II), P89 (HG9), and VEG (type III) on chrVIII, for example, we see it is clearly divided into three distinct regions (Fig. 2). For region 1, types II and III are almost identical. For region 3, types III and 9 are almost identical. Region 2 has a high number of SNPs between types II, III, and 9. Although regions 1 and 3 are consistent with the model of mating shown in Fig. 1C, region 2 is not. The SNP rate between types II and 9 is similar for all three regions of chrVIII. Across all 14 chromosomes (Fig. S1C), 42% of type III was highly similar to type II and 29% to HG9. Thus, at a low resolution it might indeed appear that type III arose from a cross between II* and β (9). However, our high-resolution whole-genome analysis shows that 29% of the type III genome is neither similar to II* nor β, establishing that the parents of type III were significantly dissimilar to the modern strains II and 9.

We similarly analyzed other strain “families” as defined by the haplotype mixture model (8): crosses A (9 × 4), B (II × 6), and D (9 × 6) (Fig. 1). Although the types I and III and HG8 genomes are largely derived from strains that are closely related to the strains in Fig. 1, 30–40% of their genomes resemble neither parental strain. This could indicate that they are not the true ancestral parents and/or that types I and III and HG8 were generated by more than one recent cross. Of note, GUY-KOE (HG5) did not have a single region in its genome that was very similar to either HG6 or HG9 (Fig. S1D), indicating that it was not derived from admixture between ancestral lineages similar to these HGs, as previously proposed (8).

From the other pairwise comparisons, we identified many genomic regions that were nearly identical between strains, suggesting they share an RCA. We thus divided the *Toxoplasma* genome into 131 haplotype blocks containing evidence of RCA between at least two strains based on boundaries identified from all 45 possible pairwise SNP plot analyses (Dataset S2A). For some haploblocks, no strain was similar (similar defined as SNP rate < 2.5 × 10⁻⁴) to any other strain, indicating that all 10 strains have distinct ancestries (“unchanged” in Dataset S2A). Our analysis focuses on haploblocks with evidence of extremely recent ancestry: pairwise SNP rates < 2.5 × 10⁻⁴, corresponding to ≥10-fold decrease in SNP rate compared with other genomic regions. Thus, our results, based on a genome-wide SNP dataset, establish that greater than six distinct ancestries are necessary to explain the relationship between these 10 HGs. Furthermore, the almost complete absence of SNPs in haploblocks shared among strains suggests that recent recombination events created most of these strains. Only the CASTELLS, GUY-KOE, and COUGAR genomes did not show evidence of RCA with any other strains.

**Construction of a Toxoplasma Haplotype Map.** The 10 *Toxoplasma* genomes provided a high-resolution SNP map that precisely defined haploblocks, but more strains were needed to draw firm conclusions about their ancestry. We therefore also identified SNPs in genome-wide RNA-sequencing data of 26 *Toxoplasma* strains representing 12 of the 14 currently defined HGs. We identified 606,742 SNPs; 132,744 of which contained biallelic sequence information for all 26 strains. These SNPs were used to identify RCA (shared haploblocks) among strains as described above. Adding the RNA-seq data increased the number of haploblocks from 131 to 157 (Dataset S2B). For example, CAST (HG7) has genomics regions showing RCA with strains from types I (RH/GT1) and HG6 (BOF/FOU/GHPT) (e.g., chrVIII, Dataset S2B and Fig. S2); TgCatBr44 (HG4) has genomic regions showing RCA with TgCatBr5 (HG8), MAS (HG4), and type II (ME49/Pru/DEG) (e.g., chrVIII, Dataset S2B and Fig. S2). B73 seems to be an F1 progeny from a cross between types II and III; chrV, VI, and VIIa are derived from type III, and the rest of its genome is from type II (Dataset S2B and Fig. S2). Similarly, B41 (HG12) seems to be an F1 progeny from a type II and RAY/WTD3 (HG12) cross, as it is comprised of haploblocks almost identical to either type II or RAY/WTD3 (Dataset S2B and Fig. S2). It also becomes clear that many strains previously defined as being in the same HG are actually highly divergent. The best examples are HG5 (GUY-KOE/GUY-MAT/RUB) and HG10 (VAND/GUY-DOS), which do not share RCA for any haploblock (note the high SNP rate in Fig. S2 and see Dataset S4A). Strains from HG4 (MAS/CASTELLS/TgCatBr44) also have distinct ancestries for many haploblocks. On the other hand, types I (GT1/RH), II (ME49/Pru/DEG), and III (VEG/Cep), HG6 (BOF/FOU/GHPT), and HG12 (RAY/WTD3 but not B41) show recent common ancestry for all 157 haploblocks, suggesting they recently descended from a common progenitor.

Many strains had very recent common ancestry for chrIa, the right end of chrXI, and parts of chrXII. In particular, chrIa is highly similar between many strains, as previously observed (8, 11). If these haploblocks were transferred from a single source to many different lineages, it suggests a selective sweep driven by genetic determinants at those loci. Therefore, it is important to identify their exact genealogy, whether they have strong fitness
effects, and if so, the genes that are responsible. One of the primary difficulties in analyzing recombination among environmental isolates is that parentage is ambiguous. Thus, we must limit the conclusions drawn from such a dataset. We anchor our analysis in the genomic regions that are dissimilar between all sequenced strains. Their low similarity suggests they are derived from strains existing independently before the recent recombinations under investigation. We therefore refer to them as “unexchanged.” Using the phylogenetic topology derived from these regions as a basal genealogy, we compare the topology of the trees derived from haploblocks containing RCA to infer the ancestry of intermediate strains. For simplicity, we refer to the “directionality of transfer” of a given genomic region, disregarding true parentage and identity. Our approach is as follows: suppose 18 hypothetical strains (A–R) fall into two groups, NA-like and SA-like (Fig. 3A). For one region of the genome (tree 2), strains F and J have very low SNP rates and the rest of the genomes (tree 1) are dissimilar. To distinguish the directionality of transfer for that locus, we construct a phylogenetic tree for these two haploblocks. If the exchanged locus were derived from a J-like ancestor, we expect F to group with J (the red branch). Alternatively, if the locus were derived from an F-like ancestor, we expect J to group with F (the green branch). For each haploblock, SNP frequencies were used to calculate the pairwise distance between strains and construct neighbor-joining phylogenetic trees (Fig. S2). For most trees, there are two major clades, one containing the NA strains (type II, HG12, and COUGAR) and the other containing all other strains, many of which are from SA. Thus, COUGAR and HG12 are sister taxa of type II (Fig. S2) and an early separation and diversification occurred between NA and SA ancestral strains. Our data recapitulate the striking biallelism for most Toxoplasma genes observed in multiple studies (9, 12).

Our basal topology approach allows us to infer directionality of transfer by comparing different haploblocks with multiple outgroups. For example, chrIb_7 (Dataset S2B) falls into the unexchanged group, and chrXI_1 represents a haploblock shared between types II and III. The strains with topological differences between these two haploblocks are underlined (Fig. 3B). One can infer that a type II-like strain was the source of region chrXI_1 for CEP and VEG (type III) because COUGAR and HG12 cluster basally to II–III. A tree consisting of merely type II, CEP, and VEG would be ambiguous with regards to directionality, but the inclusion of multiple outgroups gives us enough information to make inferences about the directionality of genetic transfer between strains. Although types I and III are NA strains, their genomes are of mixed ancestry. Portions of their genomes derive from an NA type II-like strain while other parts derive from SA strains (e.g., P89-like and BOF-like, Fig. 1). ChrIa_1 (Fig. 3C), chrIa_4 (Fig. S2), and chrXI_7 (Fig. 3D) that have RCA among many strains have HG12 and COUGAR as an outgroup and cluster most closely with type II. The fact that the unexchanged topology of type II, COUGAR, and HG12 is preserved suggests that these conserved haploblocks were most likely derived from a type II–like ancestral strain. The alternate explanation, that the unexchanged topology of type II, COUGAR, and HG12 was recreated through some other set of matings is extremely unlikely. In contrast, the chrXII haploblocks with RCA among many strains have SA strains as an outgroup (e.g., Fig. 3E, chrXII_4) and were therefore most likely derived from an SA ancestral strain. In other words, the SA strain topology was preserved following the exchange of this region, suggesting that an SA strain was the source. However, the lack of a clear outgroup for most of the SA strains precludes identification of a single ancestral strain. Similarly, chrIa_2 (Fig. 3C) and chrIa_3 (Fig. S2) have no outgroup, so no definitive ancestry can be concluded. However, because chrIa_1 and chrIa_4 are derived from a type II–like strain it is reasonable to propose that these were also derived from the same type II–like strain. We performed similar analyses to determine the ancestry of most haploblocks from the Toxoplasma haplotype map (Dataset S2B).

We further determined the recombination-independent relationship between strains by assembling the 35kb genome of the Toxoplasma apicoplast (a maternally inherited organelle), identifying polymorphisms between strains, and constructing a phylogenetic tree (Fig. 3F). These data show that VEG, GT1, and TgCatBr5 inherited their apicoplast from P89-like, ME49-like, and MAS-like strains (Fig. 1). The fact that the MAS/TgCatBr5 strains clustered close to type II, with COUGAR as an outgroup, provides some evidence that MAS not only inherited chrIa from a type II–like ancestor, but also its apicoplast, which it passed to TgCatBr5.

**Crosses That Created Most Current Strains Happened Around the Same Time.** The SNP rate in haploblocks shared between two strains is related to the time of the most recent common ancestor (TMRCA). However, without knowing the Toxoplasma mutation rate, it is difficult to translate SNP rate into years. We restricted analysis to SNPs in noncoding sequences from our DNA shotgun sequencing data, which are less likely to be subject to positive or diversifying selection and therefore evolve in a more clocklike manner. The average SNP rate for haploblocks that have RCA between different strains is low (c10^-7) but variable (Fig. 3F, Table S1), which suggests that the crosses that produced them happened relatively recently but at different times. For comparison with published TMRCA, using an intronic SNP rate of 1.94 × 10^-8 (10) and the SNP rate from our data (multiplied with a conversion factor of 2.55, which accounts for the fact that our SNP rate is lower than the real SNP rate due to incomplete genome coverage) the 6.9 × 10^-3 SNP rate for haploblocks with RCA between types I and II corresponds to a TMRCA of ~9 × 10^3 y, in concurrence with other studies (10, 11). The different SNP rates of the haploblocks shared between types II and III suggest they were transferred from a type II–like ancestor to type III two separate times (Table S1). Pairwise SNP rates for unexchanged haploblocks (Dataset S24 and Table S2) are higher than those for exchanged regions by ~60-fold. Therefore, whatever the absolute timescale, the evolutionary histories of different haploblocks clearly vary by one to two orders of magnitude.

**Identification of Highly Polymorphic Genes and Genes Undergoing Positive Selection.** Many strains had very recent common ancestry for haploblocks chrIa_1_2_3_4, chrXI_7, and chrXII_2_4 (Fig. S2 and Dataset S2B), suggesting genes in these regions might confer a selective advantage. We focused on regions chrIa_2, chrXI_7, and chrXII_2_4, as they introgressed into the most strains. Using the annotated ME49 genome (www.toxodb.org), we identified for each strain (for which we had whole-genome sequence data) which SNPs were in coding regions and if they
were synonymous (S) or nonsynonymous (NS) (Dataset S3) and selected the top five most polymorphic genes as candidates for conferring selective advantage (Table S3). For region chrXII_2_4, a promising candidate is the ROP5 cluster of pseudokinases recently shown to determine strain differences in virulence (17, 18). This cluster is present on chrXII_4, a region that is highly similar between types I, III, HG6, CAST, P89, TgCatBr44, and TgCatBr5 (Fig. 3E). Further, GT1 and VEG contain the virulent ROP5 alleles (17, 18). The chrXII_2 region, which is highly similar between 19 of the 26 strains (Fig. 3C), only contains 17 predicted genes, most of unknown function. For chrXII_7, a potential candidate is Tgd057, recently shown to contain an H-2K(b) CD8+ T-cell epitope (19).

We also identified polymorphic genes under positive selection for the whole genome. From the 45 pairwise SNP comparisons, we focused on genes with the highest average SNPs/kb and the highest NS/S rate in the greatest number of pairwise comparisons. We excluded duplicated genes, which have a high rate of false positive SNP calls. Polymorphic genes (≥10 times ranked in top 200 polymorphic genes for all 45 pairwise strain comparisons) and divergent genes (NS/S > 2) were enriched for genes encoding proteins with a signal peptide ($P = 0.0005$ and $P = 4.8 \times 10^{-7}$, respectively, hypergeometric distribution). A list of genes that are most polymorphic and divergent (and have ≥4 SNPs), have a signal peptide and are highly expressed (≥90th percentile, toxodb.org) in at least one life stage, is shown in Table S4. This table contains four rhapsody protein kinases, two of which, ROP16 and ROP18, are known to be involved in mediating strain-specific differences in virulence (20, 21). ROP17 and ROP39 have all residues important for kinase activity (22) but their function is unknown. In addition, Table S4 contains three genes: Tgd057, GRA6, and GRA4, which encode proteins with H-2K(b) (Tgd057) and H-2L(d) (GRA4 and GRA6) T-cell epitopes (23, 24). The other genes in this list should also encode proteins that are good candidates for either direct host-cell modulation or strong antigenicity.

Discussion

Here we present a comprehensive representation of global Toxoplasma diversity based on whole genome analysis. Our genome-wide SNP data were used to make a haplotype map from which we inferred the ancestry of each haplotype. It is interesting to
note that haploblocks recently derived from a type II–like strain are present in 18 of the 25 other strains for which we have genomewide SNP data. The only strains not containing type II–like haploblocks with low SNP rates were CASTELLS and the strains isolated in a Guianese rainforest GUY-KOE/GUY-MAT/GUY-DOS/VAND and RUB (7). In such a remote area, one can speculate that these Guianese strains have not had recent contact with NA type II–like strains. Previous analysis of five concatenated loci indicated that the COUGAR strain was particularly divergent compared with all others and it was estimated that the TMRCA with the other strains was ∼10 Mya (8). Our sequencing results, however, indicate that COUGAR clustered most closely with the type II strains and never clustered basally to all other strains. In fact, COUGAR and ME49 are as similar as MAS is to CASTELLS, two strains that are currently placed in the same HG. Our data do not support the model that most current strains are admixtures of four to six ancestral strains, as large regions of the genomes of all supposedly admixed strains are very different from those of the ancestral strains (Fig. 3B and Dataset S2B). This illustrates the importance of whole-genome analysis on studying evolution in sexual organisms, as the genomic location of strain-typing markers greatly affects the interpretation of evolutionary relationships.

We identified ≥10⁶ SNPs which may determine strain-specific phenotypes. Indeed, some of the most polymorphic genes we identified were recently described to be important for determining strain-specific differences in virulence (ROP18 and ROP16) (25, 26), modulation of host cell signaling (ROP16) (20), and interaction with the host adaptive immune response (GRA4, GRA6, and Tgd057) (23, 24). We expect that future whole-genome sequencing will enable association studies between HGs (or even specific SNPs) and phenotypes. With the exceptional tools available for gene manipulation in Toxoplasma, high-throughput testing of candidate genes (Tables S3 and S4) responsible for these phenotypes may soon uncover an unprecedented wealth of information about Toxoplasma proliferation and fitness.

We next propose a model to explain Toxoplasma’s sexual history and current geodiversity. Toxoplasma has an unusual global population distribution. Type II, followed by HG12, III, and I strains, are the dominant clonotypes in NA and E, whereas clonality is largely absent in SA. We address this transcontinental disparity by comparing pairwise relationships and haplotype maps. It was proposed that Toxoplasma originated in NA and was concurrently introduced to SA with Felidae members following reconnection of the Panamanian land bridge (∼2.5 Mya), after which NA and SA strains diverged (8). Our data corroborate this model: NA type II, HG12 (RAY/WTD3), and HG11 (COUGAR) mostly form a separate cluster from SA strains, except for the haploblocks that were originally derived from a type II–like strain and ended up in many SA strains (such as chrIa).

It has been suggested that the greater diversity of SA Toxoplasma strains might be related to the greater diversity of SA Felidae species (8). However, until the late Pleistocene extinction (∼11,500 y ago), a variety of Felidae species existed in NA as well. We instead propose that this massive extinction event, which eliminated ∼80% of large vertebrates in NA, including the NA cheetah, the American lion, and the saber-toothed tigers (Smilodon and Homotherium) (27), could have extirpated most Felidae members in NA, causing a severe bottleneck and the demise of most NA Toxoplasma strains except type II, HG11, and HG12. This might explain the limited diversity that is currently present in NA. Furthermore, the limited availability of its definitive host in NA might have led to adaptations in NA strains, such as their reduced virulence (8), enhanced cyst formation, and greater oral infectivity (28), to enhance propagation between intermediate hosts. Recent NA repopulation with SA Felidae migrants could have reintroduced some SA Toxoplasma strains to NA. Molecular genetic analysis of mountain lion (Puma concolor) DNA indicates that the current NA population is most likely derived from a population that migrated into SA ∼2.5 Mya and subsequently returned northward more recently, ∼10,000 y ago (29). Coincident with this remigration of SA Felidae to NA, a sudden burst of outcrossings could have occurred in which a NA type II–like strain and SA strains similar to HG9 and HG6 crossed to create types I and III, and HG6. The type I, type III, and HG6 lineages inherited the region containing the virulent ROP5 gene cluster (chrXII_4) from the SA strain, and chrIX_2 and chrXL_7 from the type II–like strain, which together might have conferred a strong selective advantage. Recent global trade and the spread of the domestic cat probably contributed to further Toxoplasma dissemination and could account for the rare SA isolates in other parts of the world, e.g., MAS, isolated in France.

Part of a type II–like chrIa is present in many SA strains, suggesting that it confers a selective advantage. Hybridization of genomic DNA of multiple strains to a microarray that contained probes to distinguish types I, II, and III was recently used to determine relative similarity of multiple strains to type I, II, or III (11). Confirming our analyses, it was reported that there are regions where F89 and VEG are highly similar and regions where GT1 and FOU (HG6) are highly similar. However, the absence of type II–like regions in SA strains was used to argue that it was unlikely that chrIa was derived from a type II–like NA strain, which is supported by our data. However, the hybridization approach has several limitations when used to identify haploblocks: (i) 53% of the genome of type II is almost identical to type II or type I (Dataset S2A) and these regions can therefore not be investigated by this approach. (ii) The similarity of nontype I, II, and III strains to each other cannot be investigated. (iii) The method has lower resolution, as there are only 1,500 usable probe sets for the whole genome. Our high-resolution genomewide SNP comparisons clearly show that some of these strains do contain type II–like regions: haploblocks VIIa_1_2 and XL_6_7 (Fig. 3D) from the HG6 strains and haploblock VIIb_10_11 and VIII_1_2_3_4 from TgCatBr44 have a type II–like ancestry (Fig. S2). The fact that part of the HG6 genome is derived from a type II–like strain provides an important clue to a potential route of transfer of an NA type II–like chrIa into many SA strains. HG6 is one of the few HGs that is found worldwide and ∼7% of its genome has RCA with the HG9 strain P89, which has been isolated from both NA and SA. Many other strains contain haploblocks with RCA with HG6 and HG9 (Fig. S2 and Dataset S2B) including type I, type III, TgCatBr44, CAST, and TgCatBr5. It is therefore possible that a relatively recent cross introduced chrIa from a type II–like strain to an ancestral HG6–like strain, which might have crossed with other ancestral HG6 strains diluting the type II–like genomic portion, and then recently crossed with an ancestral HG9 strain and so on. With each new cross, the type II portion of the genome would be diminished and only the parts with a selective advantage retained. What trait chrIa confers is currently unknown but previous studies indicate it is not enhanced virulence or oral transmission (8). The phenotype may only be detected in the definitive feline host, such as enhanced fecundity or oocyst viability. Our genome analysis refined the region of chrIa that is conserved among most strains to a region of only 17 predicted genes and feline studies may identify the fitness advantage it confers.

The model we describe combines the striking evidence of ancestry and recombination that we uncovered using whole-genome sequencing with the history of mammalian migration and extinction discerned from the fossil record. It is a model of Toxoplasma evolution that will be continually tested as more complete genomes are sequenced and we see the full impact of its sexual history.
Materials and Methods

Genomic Sequencing. T. gondii was cultured on human foreskin fibroblasts as described previously (30); all strains in this study have been described and some are available at BRC Toxoplasma (http://www.toxorb.com) (8, 30). DNA was isolated from lysed parasites using DNAzol (Invitrogen) and prepared for high-throughput sequencing as per the Illumina single-end genomic DNA kit protocol (COUGAR, CASTELL’S, and MAS). Thirty-six nucleotides of each library was sequenced on an Illumina GAII and processed using the standard Illumina pipeline. Paired-end sequencing libraries were constructed for P9B, GUY-KOE, TgCatBrS, and BOF using the Nextera Illumina-compatible DNA sample prep kit (Epigenome) and amplified with the modified PCR protocols (31). Libraries were barcoded (four strains per lane) and paired-end sequenced on Illumina HiSeq2000 (40-40 nucleotides). All libraries were spiked with trace amounts of the phiX174 bacteriophage DNA for quality control. Sequence reads were aligned to the ME49 genome using the Maq software package (32). Reference genomes from ME49, GT1, and VEG were obtained from toxodb.org (release 6.3). Genome coverage is shown in Table S5. All data presented are from reads aligned to the ME49 genome. There was no particular bias of read density or coverage for any individual chromosome or region. Data are accessible at the Short Read Archive under accession no. SRA047110.

High-Throughput RNA Sequencing. Murine bone marrow-derived macrophages (30) were infected with different T. gondii strains. After 20 h, total RNA was extracted (Qiagen RNeasy Plus Kit) and integrity, size, and concentration of RNA were checked (Agilent 2100 Bioanalyzer). The mRNA was then purified (Dynabeads mRNA Purification Kit; Invitrogen), fragmented into 200-400 base-pair-long fragments and reverse transcribed into cDNA before Illumina sequencing adapters were added to each end. Libraries were barcoded and paired-end sequenced (SRA050293). Number of reads that mapped to the ME49 genome per strain is shown in Table S6.

T. gondii Interstrain SNP Mapping. The complete GT1 and VEG genome sequences were aligned to the ME49 genome using the mummer software package and SNPs were identified using nucmer with standard settings (33). Reads for all other genomes and RNA-seq data were aligned to the ME49 genome and high-confidence SNPs were called using the easyrun option of the Maq software package (32). To construct the apicoplast genome for the sequenced strains, we aligned reads to the published (GenBank: U87145.2) complete RH (type I) apicoplast genome. The GT1, VEG, and ME49 apicoplast genomes were constructed by aligning reads from the trace archive to the RH apicoplast genome using blast (34) and assembling them using Sequencer software (Gene Codes Corporation). The location of each polymorphism was assigned based on the physical map of the ME49 genome or the RH apicoplast genome. More details are in SI Materials and Methods.

ACKNOWLEDGMENTS. We thank C. Su and M. Darde for contributing T. gondii strains; Asis Khan, David Sibley, and John Boothroyd for useful discussions; the Massachusetts Institute of Technology biomicrocenter for excellent technical assistance; The Institute for Genomics Research and the Wellcome Trust Sanger Centre, Cambridge, United Kingdom for the GT1, ME49, and VEG genome sequences; and the Toxoplasma gondii database for use of T. gondii genome data. J.P.J.S. was supported by a Massachusetts Life Sciences Center New Investigator Award, National Institutes of Health R01-A008621, and the Pew Scholars Program in the Biomedical Sciences; M.B.M. was supported by the Knights Templar Eye Foundation; W.N. and D.L. were supported by Pre-Doctoral Grant in the Biological Sciences S-T32GM007287-33; and S.M. was supported by training Grant T32AI060516.


Minot et al. PNAS | August 14, 2012 | vol. 109 | no. 33 | 13463