A Monomorphic Haplotype of Chromosome Ia Is Associated with Widespread Success in Clonal and Nonclonal Populations of *Toxoplasma gondii*

Asis Khan,[a] Natalie Miller,[b] David S. Roos,[b] J. P. Dubey,[c] Daniel Ajzenberg,[d] Marie Laure Dardé,[d] James W. Ajioka,[e] Benjamin Rosenthal,[c] and L. David Sibley[a]

Department of Molecular Microbiology, Washington University School of Medicine, St. Louis, Missouri, USA[a]; Department of Biology and Penn Genome Frontiers Institute, University of Pennsylvania, Philadelphia, Pennsylvania, USA[b]; Animal Parasitic Disease Laboratory, Animal and National Resources Institute, Agricultural Research Service, U.S. Department of Agriculture, Beltsville, Maryland, USA[c]; Centre National de Référence (CNR) Toxoplasmose/Toxoplasma Biological Resource Center (BRC), Centre Hospitalier-Universitaire Dupuytren, and Laboratoire de Parasitologie-Mycologie, EA 3174-NETEC, Faculté de Médecine, Université de Limoges, Limoges, France[d]; and Department of Pathology, University of Cambridge, Cambridge, United Kingdom[e]

ABSTRACT ***Toxoplasma gondii*** **is a common parasite of animals that also causes a zoonotic infection in humans. Previous studies have revealed a strongly clonal population structure that is shared between North America and Europe, while South American strains show greater genetic diversity and evidence of sexual recombination. The common inheritance of a monomorphic version of chromosome Ia (referred to as ChrIa*) among three clonal lineages from North America and Europe suggests that inheritance of this chromosome might underlie their recent clonal expansion. To further examine the diversity and distribution of ChrIa, we have analyzed additional strains with greater geographic diversity. Our findings reveal that the same haplotype of ChrIa* is found in the clonal lineages from North America and Europe and in older lineages in South America, where sexual recombination is more common. Although lineages from all three continents harbor the same conserved ChrIa* haplotype, strains from North America and Europe are genetically separate from those in South America, and these respective geographic regions show limited evidence of recent mixing. Genome-wide, array-based profiling of polymorphisms provided evidence for an ancestral flow from particular older southern lineages that gave rise to the clonal lineages now dominant in the north. Collectively, these data indicate that ChrIa* is widespread among nonclonal strains in South America and has more recently been associated with clonal expansion of specific lineages in North America and Europe. These findings have significant implications for the spread of genetic loci influencing transmission and virulence in pathogen populations.**

IMPORTANCE **Understanding parasite population structure is important for evaluating the potential spread of pathogenicity determinants between different geographic regions. Examining the genetic makeup of different isolates of *Toxoplasma gondii* from around the world revealed that chromosome Ia is highly homogeneous among lineages that predominate on different continents and within genomes that were otherwise quite divergent. This pattern of recent shared ancestry is highly unusual and suggests that some gene(s) found on this chromosome imparts an unusual fitness advantage that has resulted in its recent spread. Although the basis for the conservation of this particularly homogeneous chromosome is unknown, it may have implications for the transmission of infection and spread of human disease.**

Received 21 September 2011 **Accepted** 11 October 2011 **Published** 8 November 2011

Citation Khan A, et al. 2011. A monomorphic haplotype of chromosome Ia is associated with widespread success in clonal and nonclonal populations of *Toxoplasma gondii*. mBio 2(6):e00228-11. doi:10.1128/mBio.00228-11.

Editor John Boothroyd, Stanford University

Address correspondence to L. David Sibley, sibley@borcim.wustl.edu.

A.K. and N.M. contributed equally to the article.

Toxoplasma gondii is a widespread parasite of wild and domestic animals that also frequently infects humans (1). Humans are an accidental host and play no appreciable role in transmission, other than occasional infections during pregnancy that can lead to congenital toxoplasmosis (2). Human infection can result from ingestion of tissue cysts during consumption of undercooked meat from wild and domestic animals, which show high prevalence rates of chronic infection, or ingestion of spore-like stages called oocysts, which are shed in the feces of cats, thus contaminating the environment (1). Although most human infections are well controlled by healthy adults, toxoplasmosis is a frequent opportunistic infection (2) and an important cause of water-borne (3, 4) and food-borne illness (5).

Understanding the structure of *T. gondii* populations is important for evaluating the potential spread of pathogenicity determinants between different isolates, geographic regions, and hosts (6). Transmission of *T. gondii* normally occurs following sexual reproduction in the intestine of cats, resulting in shedding of diploid oocysts, which undergo meiosis in the environment (1). Despite this capacity, sexual recombination between strains of

T. gondii in the wild appears to be exceedingly rare, at least in some localities (6). Instead, the parasite can be transmitted frequently by asexual propagation, which can occur in a variety of hosts, and can spread through the food chain by carnivorous or omnivorous feeding (7). Additionally, recent studies of outbreaks have emphasized the potential for self-mating to expand a single clonal haplotype that subsequently infects many hosts due to oocyst shedding by a single infected cat (8).

The most extreme example of clonality, which may occur by self-mating, crossing of virtually identical genotypes, or asexual transmission, is the overabundance of three lineages, designated types 1, 2, and 3 (previously called I, II, and III), that comprise the majority of strains in both North America and Europe (6). Recently, a fourth clonal lineage, designated haplogroup 12, has been identified based on isolates that are common in wild animals in the United States (9). In contrast, *T. gondii* strains from South America comprise distinct genetic groups that are genetically more diverse, reflecting a greater frequency of sexual recombination (10–13). The factors that result in these very different population structures are largely unknown, but they suggest that fundamentally different modes of transmission predominate in these regions. Consistent with their different genetic structures, there is very little evidence of recent gene flow between south and north (10), although this conclusion is based on relatively few genetic loci.

Previous studies have revealed that the three clonal lineages of North America and Europe share a common recent ancestry and arose though only a few genetic recombinations in the wild (7, 14). Importantly, a progenitor of the type 2 lineage served as one of the parents in each of these crosses, the other parents being derived from distinct lineages (9, 14). These ancestral strains were themselves closely related, since most genes differ by only 2 to 3% at the DNA level (7, 14). Importantly, this recent shared ancestry led to a pattern of biallelic diversity among the three clonal lineages (14). Based on the rate of diversity of these lineages, revealed by derived polymorphisms, it is estimated that they share a recent common ancestry of within the last 10,000 years (7). Following this genetic bottleneck, the clonal lineages rapidly expanded, either widening their biological niches or replacing existing strains that were less fit.

Coincident with this recent expansion, all three clonal lineages also inherited a common version of chromosome Ia (referred to as ChrIa*), which as a consequence exhibits extremely low levels of diversity among their descendents (15). This pattern of inheritance has led to the suggestion that genes found on ChrIa* may confer some fitness advantage, although the biological basis of this trait remains uncertain. Intriguingly, some South American strains also harbor regions of ChrIa* (10), although they are otherwise genetically distinct populations. It has been suggested that this pattern might represent recent introgression of northern strains into South America (10) or alternatively that specific South American lineages spread northward (12), expanding clonally to become predominant. Genetic exchange between populations is relevant to understanding the potential spread of pathogenicity determinants, such as ROP18, a major virulence factor in the mouse model of toxoplasmosis (16–18), which shows a genetic interaction with a locus on ChrIa (18; also unpublished data). For example, ROP18 exists as both virulent and avirulent alleles that show evidence of long-term balancing selection (19), implying that these very different phenotypic versions are adaptations to distinct environmental niches.

In the present study, we have examined a wide variety of geographically diverse strains, assessing their relationships based on intron sequences presumed to be selectively neutral and genome-wide array-based genotyping, in order to provide a model for the emergence and dominance of a monomorphic version of ChrIa.

RESULTS

Population genetic structure of *T. gondii*. To determine the relatedness among different strains of *T. gondii*, we characterized 74 isolates that were obtained primarily from North America, Europe, and South America. Strains were isolated from a wide range of wild and domestic animals and also from humans (see Table S1 in the supplemental material). Additionally, several African ($n = 5$) (20) and Chinese ($n = 2$) (21) isolates were characterized using sequence-based intron markers (10), providing greater resolution for estimating genetic diversity than the restriction fragment length polymorphisms (RFLP) or microsatellites (MS) previously used to type these strains. For reference, several strains from the 12 previously defined haplogroups were also included (9, 10).

Strains were characterized by sequencing 8 intron markers from 5 unlinked genes (located on ChrIV, ChrVIIa, ChrIX, ChrX, and ChrXI), as described previously (10), collectively comprising 3,787 bp from each strain. An unrooted neighbor-net network was developed using the single nucleotide polymorphisms (SNPs) present in the intron sequences (Fig. 1A). Neighbor-net analysis showed a number of major clusters, which we defined as 14 separate haplogroups of *T. gondii* (Fig. 1A). The first 12 of these haplogroups correspond to the previously defined haplogroups of *T. gondii* (9, 10). Recently isolated strains from China and Africa were found to comprise two additional haplogroups, distinct from those previously described. Strains TgCtPRC2 and TgCtPRC6 were identical to each other and represent a common genotype in China that was previously defined by RFLP markers (21, 22); here they defined a unique haplogroup called 13 (Fig. 1A). In addition, several isolates previously typed based on MS markers as *Africa 3* (20) were clustered in a separate node defined as haplogroup 14 (strains TgA105003 to TgA105006) (Fig. 1A). Another African isolate, previously defined as *Africa 1* (20), was clustered with haplogroup 6 (strain TgA105001) (Fig. 1A), consistent with this group being found in Africa and South America

The population structure of *T. gondii* was also reconstructed using a Bayesian Monte-Carlo Markov chain sampling method that was implemented in the software program STRUCTURE (23). The ancestry among the different strains was analyzed based on polymorphisms in the intron sequences. We used an ancestral model with independent allele frequencies to estimate the number of ancestral populations (K) needed to explain the current population structure by minimizing departures from Hardy-Weinberg expectation. The ancestral population size of 6 ($K = 6$) closely resembled the population structure developed by neighbor-net analysis and was chosen as the best fit for the current data based on an *ad hoc* statistical analysis of the rate of change in K, as described previously (24) (Fig. 1B; see also Fig. S1 in the supplemental material). The pattern for lower K values led to artificial clustering of some groups that were shown to be distinct by neighbor-net analysis, while higher values provided no greater resolution (see Fig. S2). STRUCTURE analysis revealed both shared ancestry, as reflected by similar color blocks, and distinct separation between

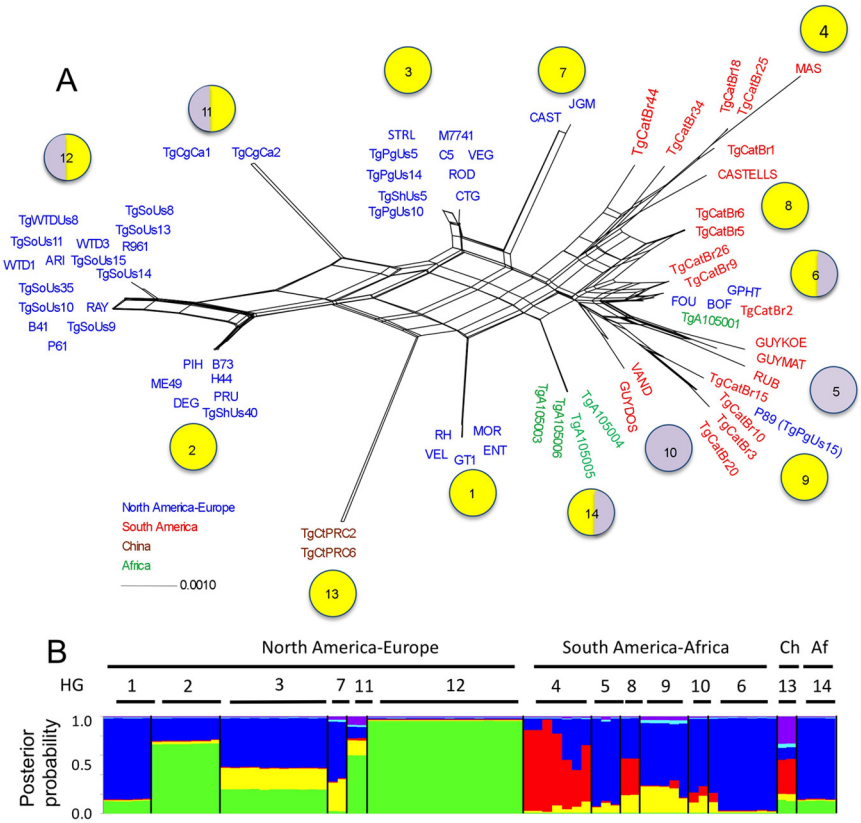

FIG 1 Population structure analysis of *T. gondii*. (A) Neighbor net developed from analysis of 5 intron sequences from 74 *T. gondii* strains. Strong geographic separation was evident between North American-European (blue lettering) and South American (red lettering) *T. gondii* strains. The major branches were classified into 14 haplogroups (numbers in circles). Circles indicate monomorphic ChrIa* (yellow) and divergent ChrIa (blue). Further details on strains are provided in Table S1 in the supplemental material. (B) STRUCTURE analysis of 74 *T. gondii* isolates, conducted using concatenated intron sequences. An ancestral population size of $K = 6$, based on a linkage model, was chosen. See Fig. S2 for additional K values. Haplogroups (HG) are listed at the top. Ch, China; Af, Africa.

North America-Europe and South America (Fig. 1B; see also Fig. S2). Interesting, haplogroups 7 and 11 in North America and haplogroup 13 from China showed greater genetic recombination than other groups, while African haplogroup 14 was closely related to haplogroups 1 and 6, which share much of their ancestry in common (Fig. 1B; see also Fig. S2).

Composition of ChrIa. Previous studies have demonstrated that three archetypal northern clonal lineages of *T. gondii* share a monomorphic version of ChrIa (a haplotype designated ChrIa*), which was likely acquired coincident with their recent common ancestry (10, 15). To determine the genetic pattern of ChrIa in newly studied isolates, we sequenced 12 reference blocks (~800 to 900 bp each) of ChrIa from the new isolates and compared them to previously studied strains (10, 15) (see Fig. S3 in the supplemental material). These 12 blocks were chosen to span the chromosome at intervals without regard to coding function (10). Network analysis of the ChrIa sequences revealed that all of the strains fall into only four major clusters (see Fig. 3B). The majority of strains contained versions of chromosome Ia that were highly similar or identical to the ChrIa* haplotype (yellow in Fig. 2). Previously described in types 1, 2, and 3, ChrIa* was also found in most strains of haplogroups 4, 7, 8, and 9 (Fig. 2; see also Fig. S3 in the supplemental material). The exception to this pattern is comprised of the highly divergent groups 5 and 10, which are characterized by completely different versions of ChrIa in each isolate (blue in Fig. 3B). Importantly, haplogroup 13 isolates from China also clustered with the ChrIa* haplotype (Fig. 2), revealing that they contain a version of this chromosome that differs from ChrIa* by only a few shared SNPs. African haplogroup 14 clustered with haplogroup 6, together comprising the 3′ chimeric group (green in Fig. 2). Group 11 together with many members of haplogroup 12 from North America forms a 5′ chimeric grouping (red in Fig. 2). Comparison of the related strains from each group suggests that each of the 3′ and 5′ chimeric groups may represent a single meiotic recombination in the wild, since members of these groups show strong conservation of their haplotypes across ChrIa (see Fig. S3). Additionally, each of these three groups (i.e., 5′ chimeric, 3′ chimeric, and purely monomorphic) shows minor variation within it, likely reflecting mutations arising since the initial founding events.

Disparate ancestry of ChrIa versus the genome as a whole. The conservation of the same version of ChrIa* among so many diverse lineages suggested that it might have an ancestry different from that of other regions of the genome. To compare their ancestries, we generated neighbor networks either from the intron sequences, representing the genome as a whole, or the sequenced blocks on ChrIa. Because we were specifically interested in the origin of ChrIa*, we included only those strains that contained the

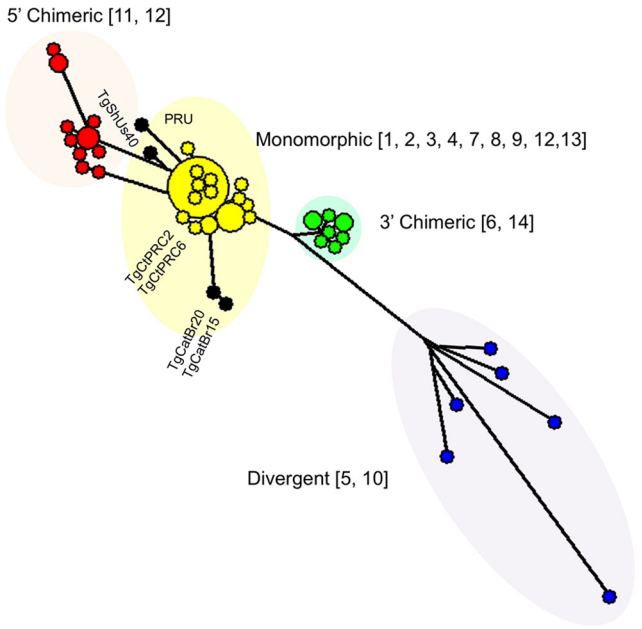

FIG 2 Network analysis of sequenced regions from ChrIa reveals four major clusters. Green indicates 3′ chimeric, red indicates 5′ chimeric, yellow indicates the monomorphic ChrIa* haplotype, and blue represents divergent ChrIa. Haplogroup numbers are indicated in brackets. Sizes of spheres are proportional to the numbers of strains. Strains included are listed in Table S1 in the supplemental material. See Fig. S3 for a diagram of regions included in the analysis.

majority of the monomorphic version of this chromosome (corresponding to the yellow cluster in Fig. 2; see also Table S1 in the supplemental material). Neighbor-net analyses of sequences from several introns showed substantial genetic divergence between northern and southern strains (Fig. 3A). This divergence is evident in two respects: first, strains from North America and Europe versus those from South America separate to the left and right portions of the network, respectively. Second, the branch lengths, representing differences between individual groups, are relatively long. In contrast, the neighbor-net of ChrIa* showed very little divergence among these lineages, as reflected by the very short branch lengths (Fig. 3B; note the different scale). The discrepancy in the distances spanned by these two respective networks argues for a very different ancestry for the genome as a whole versus ChrIa*.

Time estimates for the divergence. The relative abundance of polymorphisms among related lineages reflects the time of divergence from a common ancestry; such changes arise due to spontaneous mutation, and under an assumption that changes are neutral, the frequency of SNPs can be used to estimate the time since the most common recent ancestry (TMRCA) (25). Comparison of TMRCA estimates for the genome as a whole, based on SNPs present in the intron sequences, versus those found in regions of ChrIa*, suggests very different ancestries for these regions of the genome (see Table S2 in the supplemental material). Previous estimates from intron sequences using a wider range of strains puts the divergence of lineages between North America-Europe and South America on the order of 10^6 years (10) (Fig. 4). In order to compare our data to those in this previous report, we used similar estimates for the neutral mutation rate (see Table S2) (25, 26). When the highly divergent haplogroups 5 and 10 were excluded from analysis, an estimate of 10^5 years since common ancestry was obtained based on SNPs found in the intron sequences for differences between North American-European and South American strains (Fig. 4; see also Table S2). The age of common ancestry within South American strains (groups 4, 8, and 9) was also estimated at ~10^5 years (Fig. 4; see also Table S2). In contrast, the age of common ancestry of North American and European strains (groups 1, 2, 3, 7, and 12) was on the order of 10^4 years, similar to that previously reported (7). Importantly, the age of ChrIa* from isolates in both North America-Europe and South America was estimated to be only ~10^4 years, reflecting the fact that it lacks biallelic polymorphisms seen in other regions of the genome (Fig. 4; see also Table S2).

Previous estimates of the mutation rate presumed that the substitutions that distinguish human and primate malaria lineages have accumulated over the several million years of independent evolutionary history. Those mutations must have accumulated over a far shorter time if, as now seems well established, malaria caused by *Plasmodium falciparum* was acquired by human ancestors far more recently than previously suspected (27, 28). Independent support for a far faster mutation rate has also recently been derived from the coevolution of avian species with their respective malarial parasites (29). Using such a revised rate, the age estimates for *T. gondii* haplogroups are 10-fold lower (Table S2). In spite of uncertainty in the absolute age, it remains clear that diversity has been accumulating in ChrIa* for far less time than it has been accumulating in the genome as a whole, even allowing for possible differences in the rate at which mutations in different regions of the genome may have become fixed within the population.

Evidence for gene flow between South and North America. One attribute of many *T. gondii* populations is a high degree of linkage disequilibrium, resulting in large blocks of the genome being haplotype specific (14). Because outcrossing is relatively rare in isolates from North America and Europe, such conserved haploblocks could potentially be used to detect the presence of introgression between divergent lineages. To detect such genetic exchange between northern and southern strains, we took advantage of a photolithographic microarray for *T. gondii*, which was designed based on the ME49 (type 2) genome and which contains, in addition to expression profiling probes, a variety of probe sets designed to detect SNPs specific to each of the type 1, 2, and 3 clonal lineages (30). Because these three clonal lineages are so closely related, SNPs are relatively rare (approximately 1 in 100 bp) and typically occur as a single nucleotide change at a particular position, where the other two lineages share a second nucleotide. The pattern of inheritance of such strain-specific SNPs across the genome has been used previously to define haplotypes across major portions of the chromosomes (14). Hybridization with DNA from strains GT1 (type 1), ME49 (type 2), and CTG (type 3) indicated perfect concordance with genotype and SNPs defined by the probes (Fig. 5A). As evidence of this, all the dots in the top row of the GT1 sample were red, corresponding to a positive signal for type 1 alleles, while all of the dots in the second or third rows were gray, corresponding to negative signals for type 2 and 3 SNPs (Fig. 5A). Similar genotype-specific patterns were seen for ME49 (type 2 specific) and CTG (type 3 specific) (Fig. 5A). The lack of positive calls in some regions (e.g., ChrIV in GT1 and ME49 or ChrXI in ME49 and CTG) does not indicate poor concordance

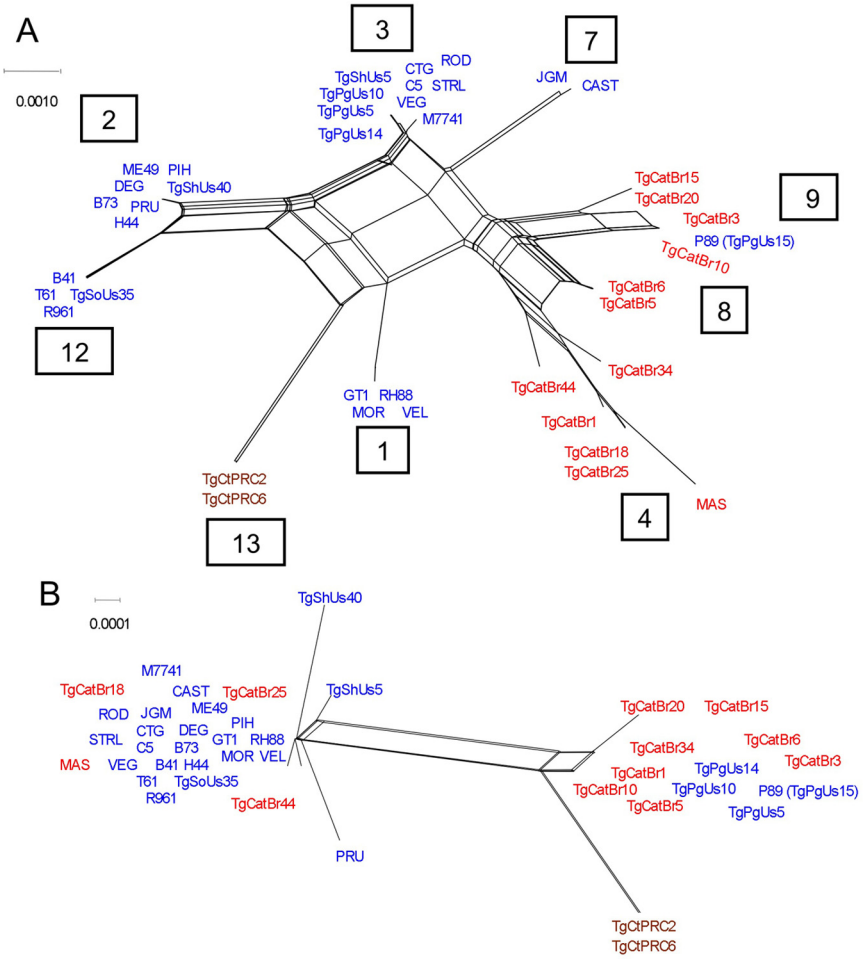

FIG 3 Neighbor network analyses of the genome as a whole based on intron sequences versus regions of ChrIa. (A) Overall genetic relatedness based on introns shows a strong split between North America-Europe and South America that is based on moderately deep branches. (B) In contrast, ChrIa* contains very low diversity among these same strains. Note that the scale differs 10-fold from that in panel A. Color coding indicates the continent of origin: North America and Europe (blue), South America (red), or Chinese (maroon). Strains included contain monomorphic ChrIa* (yellow cluster in Fig. 2; see also Table S1 in the supplemental material).

but merely the lack of typeable SNPs (i.e., ChrIV is virtually identical in GT1 and ME49, so it is not possible to definitively distinguish between these strains at these loci). ChrIa lacks significant polymorphisms defined by the expressed sequence tag (EST)-based or genetic marker probe sets and hence does not have sufficient SNP density for analysis by this method.

The widespread distribution of the ChrIa* haplotype in North America and Europe and less-extensive distribution in South America have previously been used to argue for recent spread from north to south by genetic recombination (10). Having established a platform for evaluating strain-type-specific SNPs, we hybridized a series of genomic DNA (gDNA) samples from *T. gondii* strains isolated in South America and evaluated them for haploblocks of SNPs that would be indicative of recent introgression from type 1, 2, or 3 strains in the north. The most striking pattern was the conservation of regions where type 3 (blue dots on ChrIV) or type 1 (red dots on ChrVIIb right, VIII left, XI, and XII) alleles showed a strong strain-dependent predominance in the clonal lineages; these regions were also conserved among most South American strains (Fig. 5A). Notably absent among southern lineages were haploblocks of shared ancestry with type 2 (green dots), the exception being the right end of ChrVI and left end of ChrXII among some strains (Fig. 5A and data not shown). In portions of the genome where SNP diversity makes it possible to clearly distinguish between haplogroups 1, 2, and 3, it is evident that South American strains exhibit diverse patterns of inheritance. For example, TgCatBr3 (haplogroup 4) and TgCatBr15 (haplogroup 9) share similar hybridization patterns across ChrIX, but these strains differ substantially on ChrVIIa (Fig. 5A). These patterns of inheritance likely reflect a relatively high rate of sexual recombination in South American strains. Strains harboring ChrIa* discussed above are highlighted in yellow (Fig. 5A); it is not possible to identify any particular regions of the genome whose inheritance pattern is shared with this chromosome.

To provide additional insight into the extent of shared ancestries, array-based genotype analysis was expanded to examine concordance of haplotypes across adjacent regions of the genome (Fig. 5B). Similar to the above description, colored dots were used at the top of each panel to indicate a definitive match, while dots at the bottom of each panel indicate an absence of hybridization to

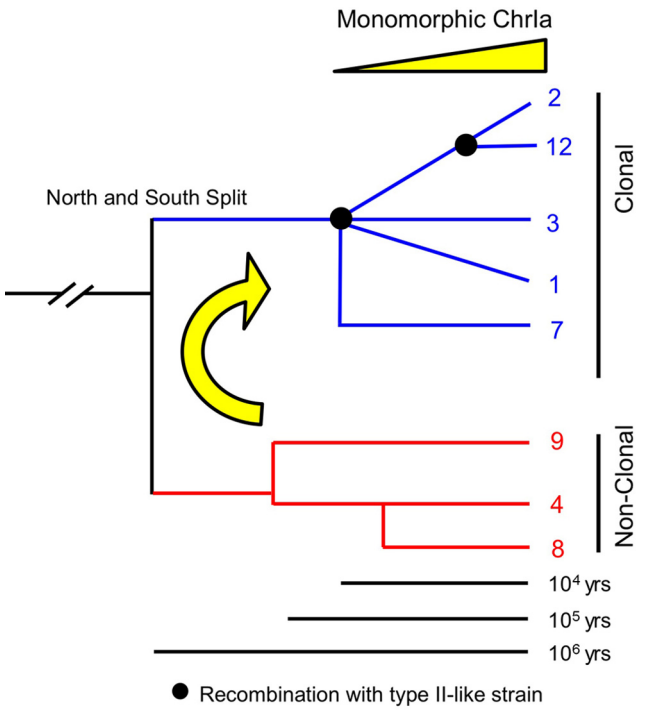

FIG 4 Model of evolutionary history of *T. gondii* lineages. The separation of *T. gondii* populations from North America and South America was estimated to have occurred 10^6 years ago. The monomorphic ChrIa* haplotype was found in the nonclonal South American haplogroups 4, 8, and 9, which have diverged over the past 10^5 years, and within the North American-European clonal lineages, which have a more recent common origin within the last 10^4 years. Age estimates are from Table S2 in the supplemental material and reference 24.

the alternative alleles (Fig. 5B). In addition, the consistency of haplotypes was plotted graphically between the rows of allelic patterns (dots) over an 11-SNP sliding window: as expected, the red line was always uppermost for GT1, the green line for ME49, and the blue line for CTG (Fig. 5B). These patterns reveal genotype information not directly apparent from individual allele-specific SNPs, based on the concordance of shared blocks, defined here as haplotype consistency. Shared ancestry can be inferred from these patterns based on the assumption that recently derived haploblocks should show a high degree of consistency across a broad region, while disruption of the consistency plot would indicate recombination or mutational drift.

Hybridization patterns for additional strains, presenting type 1-, 2-, or 3-specific SNPs in separate graphs, were superimposed on the patterns observed for the GT1, ME49, and CTG reference strains (shown in black) for comparison. By comparing the allelic patterns and extent of haplotype consistency for a variety of strains from South America, several patterns are made evident. It has been suggested that P89 (TgPgUs15) may be similar to the ancestral strain that gave rise to type 3 (14), and indeed, strong type 3 similarity was evident in many regions of the P89 (TgPgUs15) genome (i.e., ChrII), although other regions were moderately (i.e., ChrX), or highly (i.e., ChrXI) divergent (Fig. 5C). Many of the regions where P89 (TgPgUs15) differed from CTG corresponded to regions where types 2 and 3 are highly similar, suggesting they both inherited these regions from the same non-P89 parent. Although many South American strains also show large blocks of type 1 SNPs (Fig. 5A and unpublished data), the low degree of haplotype consistency among these regions suggests that they are not especially close in ancestry. The exception to this is strain FOU, a haplogroup 6 strain, which shows large blocks on chromosomes II, VI, VIIb right, VII left, IX, and XII that were highly similar to type 1 (Fig. 5D). This pattern suggests a shared ancestry of haplogroups 1 and 6, an inference also supported by STRUCTURE analysis (Fig. 1B; see also Fig. S2 in the supplemental material).

DISCUSSION

Examination of a wider array of isolates of *T. gondii* from different geographic regions uncovered a larger number of haplogroups than had been previously recognized. Additional sampling also established that the same haplotype of ChrIa (ChrIa*), initially known from clonal strains in North America and Europe, is widespread in lineages that predominate in South America, including haplogroups 4, 8, and 9, which propagate primarily sexually and do not display clonality. Despite the highly similar nature of ChrIa* in these regions, variation in other chromosomes is strongly partitioned geographically, as indicated by conserved SNP patterns that are associated with specific geographic regions. This contrast suggests that the ChrIa* haplotype has recently been transferred between comparatively diverse populations in South America into the founding stocks that subsequently expanded clonally in North America and Europe. The strong conservation of ChrIa* among divergent lineages from North America, Europe, and South America suggests that it confers a selective advantage under both asexual and sexual modes of propagation. Whether ChrIa* favors clonality in the north due to some other genetic element that is lacking in South America is uncertain; however, this possibility could be investigated by experimental genetic crosses.

In addition to the previously defined 12 haplogroups of *T. gondii* (9, 10), we have identified two new haplogroups that appear common in regions of Africa and China. Previous studies from Africa have found a mixture of canonical genotypes, including examples of types 2 and 3 that are normally found in North America and Europe (31, 32). These studies were based on sampling of domestic fowl from several countries (31, 32) and hence may have been influenced by importation of strains along with domestic animals or the relative insensitivity of RFLP markers for discovery of new polymorphisms. Studies of *T. gondii* isolates from free-range animals in Gabon, Africa, have also reported strains that are highly similar to type 3 strains, as well as those that are more diverse based on MS markers (20). Comparison of these isolates here revealed that *Africa 1* is similar to haplogroup 6. Strains with similar genotypes have previously been isolated in South America and Europe, although the latter cases are thought to have originated in Africa either through travel, immigration, or importation of contaminated meat (20, 33). The remaining *Africa 3* strains defined a new haplogroup, 14, which is distinct from those seen previously. Likewise, two Chinese strains that had previously been defined by RFLP markers (21, 22) defined a new haplogroup, called 13. More-complete analyses of their distribution and population structure (9) await wider sampling of strains from these regions.

Neighbor network and STRUCTURE analyses demonstrated a strong geographic separation of *T. gondii* strain types between

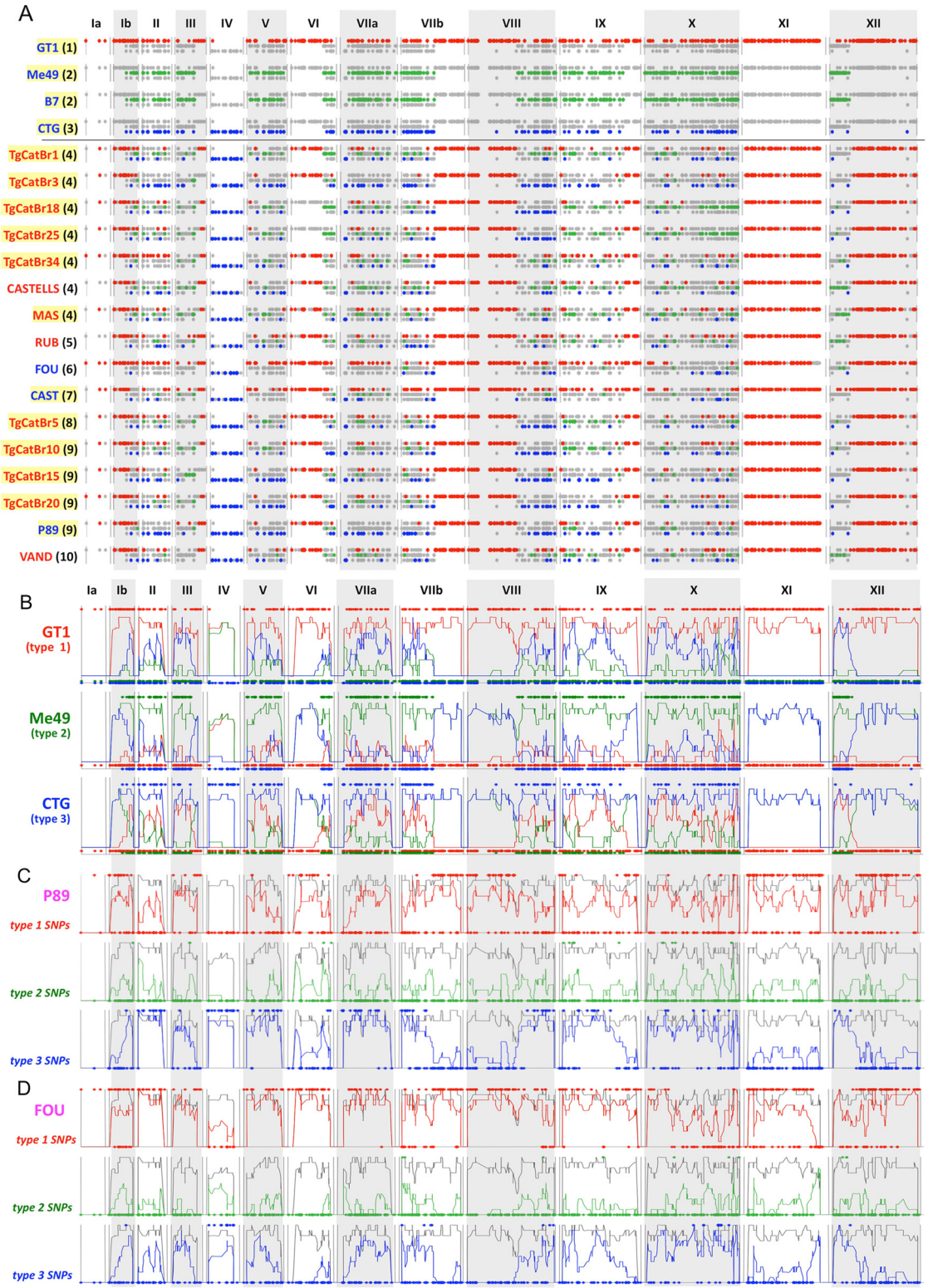

FIG 5 Genome-wide SNP typing of *T. gondii* strains based on microarray hybridization. (A) Chromosomes are indicated by alternating gray and white bars, starting with Ia on the left. For each strain, SNPs represented on the array are shown as dots, segregated into three rows representing SNPs unique to type 1, type 2, or type 3. Color indicates hybridization to the defining allele (red, type 1; green, type 2; blue, type 3), while gray indicates hybridization to the alternative allele (or failure to hybridize). Strain names at left are colored blue/red to indicate North America-European versus South American origin and shaded yellow to indicate the presence of monomorphic ChrIa*. Haplogroups are indicated in parentheses. (B) Microarray-based genotyping of *T. gondii* strains. Allelic patterns for representative type 1 (GT1), 2 (ME49), and 3 (CTG) strains. Results for P89 (C) or FOU (D) are shown based on similarity to type 1 (red), type 2 (green), or type 3 (blue). Dots above and below each graph represent strain-dependent match (top) or mismatch (bottom) hybridizations. Line graphs indicate concordance with type 1, 2, or 3 over an 11-SNP sliding window, shown as % identity on the *y* axis.

North America-Europe and South America, similar to findings in previous reports (12, 19). However, there are clearly four major ancestral populations reflecting the dominant color blocks in STRUCTURE. Based on shared ancestry, haplogroup 12 is the most likely parent that led to current-day type 2, since these two lineages share large portions of their genomes. Large haploblocks of type 2 are also seen in types 1 and 3, which likely reflects their recent admixture, as suggested previously (14). STRUCTURE also provided evidence of ancestral gene flow between north and south, as is apparent from the major color patterns that are shared across these regions at various *K* values (Fig. 1B; see also Fig. S2 in the supplemental material). Such exchanges are likely to be ancestral, because other studies have indicated that there is relatively little recent genetic exchange between these regions based on F_{ST} analysis, phylogenic analysis, and principal-component analysis (19). Additionally, strains types in each region are characterized by derived SNPs that differ from ancestral alleles, which are shared within specific northern or southern haplogroups (19). These two opposing patterns of current geographic segregation versus shared ancestry are not incompatible but rather may reflect differences in genetic exchange over time.

Despite evidence for geographic separation between different continents, all of the northern lineages and a large number of southern ones share a nearly identical haplotype of ChrIa*. In particular, nearly all members of haplogroups 4, 8, and 9 from South America share this ChrIa* haplotype despite having diversified in other parts of their genomes. The abundance of ChrIa* in South American lineages was unexpected, since the ChrIa* haplotype has previously been associated primarily with the clonal lineages 1, 2, and 3 (15). Depending on the estimate of mutational rate, the common origin of ChrIa* among these lineages is estimated to be 1,000 to 10,000 years. During this time period, extensive migration and interchange between people, livestock, rodent pests, and domestic cats may have profoundly influenced the distribution and abundance of *T. gondii*, as suggested previously (34). The apparent coincidence of the fixation of ChrIa* with the origin of the major clonal lineages led to the previous hypothesis that ChrIa* may contain genes that favor clonal transmission, driving their expansion (15). However, based on the new finding that clonal and nonclonal strains share this same version of ChrIa*, it is now apparent that the ChrIa* haplotype is associated with highly successful (i.e., abundant) strains of *T. gondii* that propagate by both sexual and asexual means.

Several possible factors could account for the paucity of diversity in ChrIa compared to the genome as a whole. It is conceivable that the ChrIa* haplotype has undergone especially strong purifying selection, enabling it to persist in an unmodified form for longer than its current diversity would otherwise indicate. Although some gene(s) found there might be subject to strong functional constraint, it is difficult to conceive how substitutions could be kept from accumulating over the entirety of the chromosome. This pattern of inheritance is unlikely to be due to drug pressure, since the majority of isolates studied here came from infected animals that were not under therapeutic treatment. Moreover, prophylactic drugs are not routinely used in humans, nor is drug resistance a common trait in *T. gondii*. However, it is conceivable that ChrIa* contains some element that results in segregation distortion or meiotic drive (35), allowing it to outcompete variant versions of ChrIa in natural crosses. Among the strains that show divergent examples of ChrIa are strains isolated from humans infected in the jungles of French Guiana (groups 5 and 10). This pattern suggests that the conservation of ChrIa* may reflect an adaptation to domestic animals or transmission by domestic cats and rodents in a cycle that is associated with human activity versus a purely sylvatic cycle. Regardless of the mechanism of its maintenance, it seems likely that the ChrIa* haplotype has recently spread via meiotic recombination between different lineages in distinct geographic regions.

Because the three northern clonal lineages (i.e., 1, 2, and 3) are thought to have arisen by genetic recombination with a progenitor of type 2 that was closely related to haplogroup 12 (9), they share large haploblocks in common that serve as a signature for recent introgression. To test for the presence of large haploblocks of type 2-specific regions within South American strains, we examined SNP hybridization patterns for contiguous stretches of clonal alleles among southern strains. Very limited evidence was found for the presence of type 2-like haploblocks in strains isolated in South America, despite the presence of a scattering of type 2-like alleles (Fig. 5). When this is combined with the fact that South American strains appear to be older, there is limited support for the model that ChrIa* arrived there by introgression from northern strains. In contrast, large blocks of type 3 SNPs were shared in haplogroup 9 strains, such as P89 (TgPgUs15), and many of these showed high haplotype consistency. A similar relationship is seen between haplogroup 6 strain FOU and the type 1 lineage. Although not an exact match for the respective parental strains, these patterns are consistent with a relatively recent ancestry of haplogroup 9 leading to haplogroup 3 and, separately, haplogroup 6 leading to haplogroup 1. Such relationships have been suggested previously based on limited genetic loci (10), and this pattern is more clearly seen in the genome-wide SNP analysis provided here. Because haplogroup 9 contains the entire monomorphic ChrIa, while haplogroup 6 is a hybrid (see Fig. S3 in the supplemental material), at present the most likely scenario is that a strain related to type 9 provided the ancestral source of ChrIa* that subsequently became widespread in North America and Europe.

Our findings reveal that a larger number of major haplotypes exist for *T. gondii* than previously recognized and yet many of these groups show strong geographic segregation. Although we have sampled a wider range of geographic regions, there remain large regions of the world that are still not well represented by current surveys of genetic diversity. Hence, it is likely that further sampling will discover additional major haplotypes. Remarkably, among the existing 14 haplogroups, 8 share a common haplotype of ChrIa*, which lacks significant polymorphism despite extensive divergence in the rest of the genome, and an additional 5 groups share large parts of this monomorphic ChrIa. The predominance of ChrIa* is paralleled by a relatively few meiotic events in the wild that gave rise to the major lineages defined here. Although the attributes that favor strains harboring this monomorphic ChrIa* are uncertain, it is associated with highly successful lineages that propagate by both clonal and sexual transmission, suggesting it imparts a general fitness advantage. Further defining of the dominance of this monomorphic ChrIa* will be informed by more-extensive population surveys, as well as experimental crosses to test its behavior in meiosis.

MATERIALS AND METHODS

***T. gondii* strains.** Strains were grown in monolayers of human foreskin fibroblast cells propagated in Dulbecco's modified Eagle's medium

(DMEM) (Invitrogen, Carlsbad, CA) supplemented with 10% fetal bovine serum, 2 mM glutamine, 20 mM HEPES (pH 7.5), and 10 μg/ml gentamicin (7, 10). Parasites were harvested after natural egress by passing through 3.0-μm polycarbonate filters to remove host cell debris (7, 10) and resuspended in phosphate-buffered saline at a concentration of approximately 10^6 cell/ml. To prepare lysates for PCR, parasites were digested with 10 μg/ml proteinase K (Sigma, St. Louis, MO) at 55°C for 2 h and heat inactivated at 95°C for 15 min (10).

PCR amplification and sequencing. Lysates were used as template DNA for PCR amplification of 5 introns (i.e., *UPRT*, *MIC*, *BTUB*, *HP*, and *EF*), as described previously (10). Amplified PCR products were sequenced using BigDye cycle sequencing (Applied Biosystems, Foster City, CA), performed by SeqWright DNA Technology Services (Houston, TX), as described previously (10).

DNA polymorphism analysis. The ClustalW/X software program was used to align the sequences using default settings. The Nexus file of the aligned sequences was imported into the SplitsTree4 (v4.11.3) software program (36) for computing an unrooted network analysis using the neighbor-net method with 1,000 bootstrap replicates.

Population structure analysis. The population structure of *T. gondii* was modeled using a Bayesian clustering algorithm implemented in the software program STRUCTURE 2.2 (23, 37) to analyze intron sequence data, as described previously (10). Twenty replicate simulations were conducted for each value of K (the number of founding groups), ranging from 2 to 10, using 10^4 burn-in repetitions and a final run of 10^4 Markov-chain Monte Carlo steps under the ancestral linkage model with independent allele frequencies. To calculate the true number of groups, K, we utilized an *ad hoc* statistical analysis based on the second-order rate of change of the likelihood function with respect to K (delta K), as described previously (24) using the software program Structure Harvester v0.6.1 (38). In short, the mean likelihood $L(K)$ (equal to Ln P D′ in STRUCTURE) was plotted based on 20 runs for each K value. Then, the mean difference was calculated between successive likelihood values of K, $L'(K) = L(K) - L(K-1)$. In the third step, we calculated the difference between successive values of $L'(K)$, where $|L''(\mathrm{K})| = |L'(K+1) - L'(K)|$. In the final step, we calculated the delta K (ΔK), $\Delta K = \mathrm{m}|L''(\mathrm{K})|/s[L(K)]$, where m is the mean of the absolute values of $L''(K)$ and s is the standard deviation of $L(K)$.

Sequence analysis of ChrIa. Twelve scattered blocks comprising 800 to 900 bp (each) were selected for sequencing as representative of ChrIa. These regions were amplified using genome-specific primers and sequenced by using the BigDye cycle, as described previously (10, 15). Sequenced regions were separately aligned using ClustalW/X and used to generate a neighbor-joining phylogeny that was viewed in the TreeView program (39). Monomorphic, shared, and divergent alleles were defined based on grouping at distinct nodes. Alternatively, network analysis was conducted using a median-joining algorithm (40) ($\varepsilon = 0$) as implemented in the software program NETWORK 4.1fluxus-engineering. ChrIa sequences from all 74 strains were aligned using the software program DNA Alignment vl.1.2.1 (fluxux-engineering), and the output file was imported into NETWORK 4.1 to develop a network using Kruskal's algorithm for finding minimum spanning trees and Farris's maximum-parsimony heuristic algorithm.

Comparison of evolution rates between introns and ChrIa sequences. ClustalW/X (41) was used to align the sequences using default settings. The Nexus file of the aligned sequences was imported into the software program SplitsTree4 v4.11.3 (36) for computing an unrooted network analysis using the neighbor-net method with 1,000 bootstrap replicates.

Age calculations. The time to most recent common ancestry (TMRCA) was estimated based on polymorphisms present in intron sequences or regions of ChrIa. Because we were specifically interested in the origin and transmission of monomorphic ChrIa among North and South American isolates, we included only those strains that contained the monomorphic version of this chromosome and that were isolated in North America and Europe versus those isolated in South America. Single nucleotide polymorphisms, i.e., those found only in a single strain, were used for estimating the time since the origin of the clonal lineages, whereas ancestral biallelic polymorphisms, i.e., those that define shared lineage-specific differences, were used to estimate the common ancestry between the north-and-south split (7, 10, 19). The formula $t = S/(\mu_a \Sigma n_i l_i + \mu_b \Sigma n_i m_i)$ was used to calculate the TMRCA, where n is the number of lineages examined at the ith locus, l_i and m_i are the number of 4-fold degeneration sites, S is the number of polymorphisms, and μ_a and μ_b are the neutral mutation rates. Estimates of μ from the closely related parasite *Plasmodium falciparum* were used for calculating the TMRCA, as described previously (25, 26).

Genotyping by hybridization to Affymetrix arrays. Genomic DNA was isolated from *T. gondii* strains grown as tachyzoites in culture and labeled for array-based genotyping as described previously (30). Briefly, DNA was isolated from cultured parasites using the Generation Capture column kit (Qiagen), and 1 μg DNA was sheared in an Invitrogen nebulizer using compressed nitrogen at 40 lb/in² for 3 min. Fragmented DNA was precipitated and labeled using the Invitrogen BioPrime array CGH genomic labeling module with biotin-14-dCTP. Labeled DNA was hybridized to the custom Affymetrix *T. gondii* microarray (http://ToxoDB.org) according to recommended protocols (30).

The *T. gondii* microarray includes several probe sets suitable for genotyping at various levels of resolution (30). Briefly, 228 polymorphisms may be interrogated using the standard Affymetrix protocol involving 40 probes per SNP (5 probes overlapping each allele, on each strand). A further 3,490 polymorphisms may be interrogated using 4 probes per SNP (one probe overlapping each allele, on each strand). Since previous results indicate comparable performance (30), these probe sets were pooled for the purposes of the present study. Hybridization to multiple replicates, including clonal cross progeny, indicated that 1,741 probe sets pass stringent performance criteria (false discovery rate, $<10^{-3}$). To improve accuracy, an additional requirement was imposed to require accurate identification of both alleles among known strains (i.e., types 1, 2, and 3), reducing the number of reliably typeable probe sets to 1,517. Genotyping calls were made using custom R scripts (30), modified to include SNPs defining all three strains for which complete genome sequence is available (type 1, GT1; type 2, ME49; type 3, VEG). Following previously described nomenclature (14), polymorphisms were defined as type 1, 2, or 3 SNPs based on the strain-defining allele (i.e., type 1 SNPs are represented by one allele in GT1 and the alternative allele in both ME49 and VEG parasites). SNP positions were mapped against the concatenated *T. gondii* genome and color coded by SNP type. The percentage of SNPs matching each type was also assessed over an 11-SNP sliding window, although other window sizes yield comparable results (not shown).

Microarray data accession number. Hybridization data are available in ToxoDB (http://toxodb.org/toxo/).

ACKNOWLEDGMENTS

We thank Chunlei Su for unpublished data and helpful discussions, Jon Boyle and Xinzhuan Su for critical review, the BRC Toxoplasma network for providing some strains, and Jennifer Barks for technical assistance.

This work was supported by a grant from the NIH (AI059176 to L.D.S.).

SUPPLEMENTAL MATERIAL

Supplemental material for this article may be found at http://mbio.asm.org/lookup/suppl/doi:10.1128/mBio.00228-11/-/DCSupplemental.

Figure S1, TIF file, 0.2 MB.
Figure S2, TIF file, 0.8 MB.
Figure S3, TIF file, 1.5 MB.
Table S1, PDF file, 0.1 MB.
Table S2, PDF file, 0.1 MB.

REFERENCES

1. **Dubey JP.** 2010. Toxoplasmosis of animals and humans. CRC Press, Boca Raton, FL.

2. **Joynson DH, Wreghitt TJ.** 2001. Toxoplasmosis: a comprehensive clinical guide. Cambridge University Press, Cambridge, United Kingdom.
3. **Bahia-Oliveira LM, et al.** 2003. Highly endemic, waterborne toxoplasmosis in North Rio de Janeiro state, Brazil. Emerg. Infect. Dis. **9**:55–62. PubMed.
4. **Benenson MW, Takafuji ET, Lemon SM, Greenup RL, Sulzer AJ.** 1982. Oocyst-transmitted toxoplasmosis associated with ingestion of contaminated water. N. Engl. J. Med. **307**:666–669.
5. **Mead PS, et al.** 1999. Food-related illness and death in the United States. Emerg. Infect. Dis. **5**:607–625.
6. **Sibley LD, Ajioka JW.** 2008. Population structure of *Toxoplasma gondii*: clonal expansion driven by infrequent recombination and selective sweeps. Annu. Rev. Microbiol. **62**:329–351.
7. **Su C, et al.** 2003. Recent expansion of Toxoplasma through enhanced oral transmission. Science **299**:414–416.
8. **Wendte JM, et al.** 2010. Self-mating in the definitive host potentiates clonal outbreaks of the apicomplexan parasites *Sarcocystis neurona* and *Toxoplasma gondii*. PLoS Genet. **6**:e1001261.
9. **Khan A, et al.** 2011. Genetic analyses of atypical *Toxoplasma gondii* strains reveal a fourth clonal lineage in North America. Int. J. Parasitol. **41**: 645–655.
10. **Khan A, et al.** 2007. Recent transcontinental sweep of *Toxoplasma gondii* driven by a single monomorphic chromosome. Proc. Natl. Acad. Sci. U.S.A. **104**:14872–14877.
11. **Khan A, et al.** 2006. Genetic divergence of *Toxoplasma gondii* strains associated with ocular toxoplasmosis, Brazil. Emerg. Infect. Dis. **12**: 942–949.
12. **Lehmann T, Marcet PL, Graham DH, Dahl ER, Dubey JP.** 2006. Globalization and the population structure of *Toxoplasma gondii*. Proc. Natl. Acad. Sci. U. S. A. **103**:11423–11428.
13. **Pena HF, Gennari SM, Dubey JP, Su C.** 2008. Population structure and mouse-virulence of *Toxoplasma gondii* in Brazil. Int. J. Parasitol. **38**: 561–569.
14. **Boyle JP, et al.** 2006. Just one cross appears capable of dramatically altering the population biology of a eukaryotic pathogen like *Toxoplasma gondii*. Proc. Natl. Acad. Sci. U. S. A. **103**:10514–10519.
15. **Khan A, et al.** 2006. Common inheritance of chromosome Ia associated with clonal expansion of *Toxoplasma gondii*. Genome Res. **16**:1119–1125.
16. **Fentress SJ, et al.** 2010. Phosphorylation of immunity-related GTPases by a parasite secretory kinase promotes macrophage survival and virulence. Cell Host Microbe **16**:484–495.
17. **Saeij JPJ, et al.** 2006. Polymorphic secreted kinases are key virulence factors in toxoplasmosis. Science **314**:1780–1783.
18. **Taylor S, et al.** 2006. A secreted serine–threonine kinase determines virulence in the eukaryotic pathogen *Toxoplasma gondii*. Science **314**: 1776–1780.
19. **Khan A, Taylor S, Ajioka JW, Rosenthal BM, Sibley LD.** 2009. Selection at a single locus leads to widespread expansion of *Toxoplasma gondii* lineages that are virulent in mice. PLoS Genet. **5**:e1000404.
20. **Mercier A, et al.** 2010. Additional haplogroups of *Toxoplasma gondii* out of Africa: population structure and mouse-virulence of strains from Gabon. PLoS Negl. Trop. Dis. **4**:e876.
21. **Dubey JP, et al.** 2007. Genetic and biologic characterization of *Toxoplasma gondii* isolates of cats from China. Vet. Parasitol. **145**:352–356.
22. **Zhou P, et al.** 2010. Genetic characterization of *Toxoplasma gondii* isolates from pigs in China. J. Parasitol. **96**:1027–1029.
23. **Falush D, Stephens M, Pritchard JK.** 2003. Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies. Genetics **164**:1567–1587.
24. **Evanno G, Regnaut S, Goudet J.** 2005. Detecting the number of clusters of individuals using the software STRUCTURE: a simulation study. Mol. Ecol. **14**:2611–2620.
25. **Rich SM, Licht MC, Hudson RR, Ayala FJ.** 1998. Malaria's eve: evidence for a recent population bottleneck throughout the world populations of *Plasmodium falciparum*. Proc. Natl. Acad. Sci. U. S. A. **95**:4425–4430.
26. **Hughes AL, Verra F.** 2001. A very large long-term effective population size in the virulent human malaria parasite *Plasmodium falciparum*. Proc. Biol. Sci. **268**:1855–1860.
27. **Liu W, et al.** 2010. Origin of the human malaria parasite *Plasmodium falciparum* in gorillas. Nature **467**:420–425.
28. **Rich SM, et al.** 2009. The origin of malignant malaria. Proc. Natl. Acad. Sci. U. S. A. **106**:14902–14907.
29. **Ricklefs RE, Outlaw DC.** 2010. A molecular clock for malaria parasites. Science **329**:226–229.
30. **Bahl A, et al.** 2010. A novel multifunctional oligonucleotide microarray for *Toxoplasma gondii*. BMC Genomics **11**:603.
31. **Bontell IL, et al.** 2009. Whole genome sequencing of a natural recombinant *Toxoplasma gondii* strain reveals chromosome sorting and local allelic variants. Genome Biol. **10**:R53.
32. **Velmurugan GV, Dubey JP, Su C.** 2008. Genotyping studies of *Toxoplasma gondii* isolates from Africa revealed that the archetypal clonal lineages predominate as in North America and Europe. Vet. Parasitol. **155**: 314–318.
33. **Ajzenberg D, et al.** 2009. Genotype of 88 *Toxoplasma gondii* isolates associated with toxoplasmosis in immunocompromised patients and correlation with clinical findings. J. Infect. Dis. **199**:1155–1167.
34. **Rosenthal BM.** 2009. How has agriculture influenced the geography and genetics of animal parasites? Trends Parasitol. **25**:67–70.
35. **Hurst GD, Werren JH.** 2001. The role of selfish genetic elements in eukaryotic evolution. Nat. Rev. Genet. **2**:597–606.
36. **Huson DH, Bryant D.** 2006. Application of phylogenetic networks in evolutionary studies. Mol. Biol. Evol. **23**:254–267.
37. **Falush D, et al.** 2003. Traces of human migrations in *Helicobacter pylori* populations. Science **299**:1582–1585.
38. **Earl DA, vonHoldt BM.** 2011. Structure Harvester: a website and program for visualizing structure output and implementing the Evanno method. Conservation Genetics Resources DOI: 10.1007/s12686-011-9548-7. http://taylor0.biology.ucla.edu/structureHarvester/.
39. **Page RDM.** 2002. Visualizing phylogenetic trees using TreeView. Curr. Protoc. Bioinformatics **2002**:Chapter 6, Unit 6.2.
40. **Bandelt HJ, Fortser P, Röhl A.** 1999. Median-joining networks for inferring intraspecific phylogenies. Mol. Biol. Evol. **16**:37–48.
41. **Higgins DG, Thompson JD, Gibson TJ.** 1996. Using CLUSTAL for multiple sequence alignments. Methods Enzymol. **266**:382–402.