Globally diverse *Toxoplasma gondii* isolates comprise six major clades originating from a small number of distinct ancestral lineages

Chunlei Su[a,1], Asis Khan[b,1], Peng Zhou[c,2], Debashree Majumdar[a,2], Daniel Ajzenberg[d,e,2], Marie-Laure Dardé[d,e], Xing-Quan Zhu[f], James W. Ajioka[g], Benjamin M. Rosenthal[h], Jitender P. Dubey[h,3], and L. David Sibley[b,3]

[a]Department of Microbiology, University of Tennessee, Knoxville, TN 37996; [b]Department of Molecular Microbiology, Washington University School of Medicine, St. Louis, MO 63110; [c]Key Laboratory of Animal Parasitology of Ministry of Agriculture, Shanghai Veterinary Research Institute, Chinese Academy of Agricultural Sciences, Shanghai 200241, People's Republic of China; [d]Centre National de Référence Toxoplasmose/*Toxoplasma* Biological Resource Center, Centre Hospitalier-Universitaire Dupuytren, 87042 Limoges, France; [e]Institut National de la Santé et de la Recherche Médicale, Unité Mixte de Recherche 1094, Neuroépidémiologie Tropicale, Laboratoire de Parasitologie-Mycologie, Faculté de Médecine, Université de Limoges, 87025 Limoges, France; [f]State Key Laboratory of Veterinary Etiological Biology, Key Laboratory of Veterinary Parasitology of Gansu Province, Lanzhou Veterinary Research Institute, Chinese Academy of Agricultural Sciences, Lanzhou, Gansu Province 730046, People's Republic of China; [g]Department of Pathology, University of Cambridge, Cambridge CB2 1QP, United Kingdom; and [h]Animal Parasitic Disease Laboratory, Animal and Natural Resources Institute, Agricultural Research Service, US Department of Agriculture, Beltsville, MD 20705

Contributed by Jitender P. Dubey, February 24, 2012 (sent for review December 6, 2011)

Marked phenotypic variation characterizes isolates of *Toxoplasma gondii,* a ubiquitous zoonotic parasite that serves as an important experimental model for studying apicomplexan parasites. Progress in identifying the heritable basis for clinically and epidemiologically significant differences requires a robust system for describing and interpreting evolutionary subdivisions in this prevalent pathogen. To develop such a system, we have examined more than 950 isolates collected from around the world and genotyped them using three independent sets of polymorphic DNA markers, sampling 30 loci distributed across all nuclear chromosomes as well as the plastid genome. Our studies reveal a biphasic pattern consisting of regions in the Northern Hemisphere where a few, highly clonal and abundant lineages predominate; elsewhere, and especially in portions of South America are characterized by a diverse assemblage of less common genotypes that show greater evidence of recombination. Clustering methods were used to organize the marked genetic diversity of 138 unique genotypes into 15 haplogroups that collectively define six major clades. Analysis of gene flow indicates that a small number of ancestral lineages gave rise to the existing diversity through a process of limited admixture. Identification of reference strains for these major groups should facilitate future studies on comparative genomics and identification of genes that control important biological phenotypes including pathogenesis and transmission.

population structure | inheritance | toxoplasmosis

The parasite *Toxoplasma gondii* commonly infects warm-blooded vertebrates and also causes zoonotic disease in humans (1). *T. gondii* has a global distribution and it infects many mammals and birds, yet all isolates of the genus *Toxoplasma* have been classified a single species (2). *T. gondii* was first isolated in the early 1900s from an African rodent (i.e., *Ctenodactylus gundi*), from which the species name was derived (3). At almost the same time, it was independently isolated from an infected rabbit in South America (4). Subsequent surveys have found *T. gondii* to be highly prevalent among many species of mammals and birds (1). Although isolates of *T. gondii* were historically considered to be highly similar, molecular analysis revealed that they display very marked clonality, notably in North America and Europe, where three predominant lineages, known as types 1, 2, and 3, comprise the vast majority of isolates (5). A fourth clonal lineage, referred to as type 12, has recently been described in North America where it is commonly found in wildlife (6). All four clonal lineages show evidence of overly abundant, highly similar multilocus genotypes, high levels of linkage disequilibrium, and only infrequent recombination. Despite their extant differences, these four lineages likely arose from a few, recent genetic crosses that occurred between a type 2 parental strain and one of several other ancestors (6, 7). Following a genetic bottleneck, these clonal lineages have rapidly expanded their ranges in the past 10,000 y (8). In contrast, an entirely distinct pattern is seen in South America, which is populated by different lineages that show markedly greater diversity within and divergence between groups (9, 10). These opposing patterns suggest that *T. gondii* propagates largely clonally in North America and Europe, but shows greater evidence of sexual recombination in South America (5). These two regions show historical patterns of interbreeding and yet currently maintain strong geographic separation (5).

The complex life cycle of *T. gondii* facilitates both clonal and sexual modes of transmission. Although many vertebrates serve as intermediate hosts for replication of haploid tissue stages, sexual development is restricted to the intestinal epithelium of cats, which shed diploid oocysts that undergo meiosis in the environment (1). Environmentally resistant oocysts are responsible for infecting herbivores and can contaminate food and water (1). Unlike related species of parasites, *T. gondii* can also be passed directly between intermediate hosts via ingestion of haploid tissue cysts during omnivorous or carnivorous feeding (8). Humans can become infected by ingesting oocysts in contaminated water or food (11, 12) or tissue cysts found in undercooked meat (13). Clonal propagation results from asexual transmission among intermediate hosts or from self-fertilization in cats encountering only one, genetically homogeneous strain.

Several methods have been developed for genotyping *T. gondii*, including restriction fragment length polymorphism (RFLP) markers that take advantage of the biallelic polymorphism displayed by the northern clonal lineages (14). Additionally, microsatellite (MS) markers have been developed to provide finer

Author contributions: C.S., A.K., D.A., and L.D.S. designed research; C.S., A.K., P.Z., D.M., and D.A. performed research; D.A., M.-L.D., X.-Q.Z., and J.P.D. contributed new reagents/analytic tools; C.S., A.K., J.W.A., B.R., and L.D.S. analyzed data; and C.S., A.K., D.A., M.-L.D., B.R., J.P.D., and L.D.S. wrote the paper.

The authors declare no conflict of interest.

Data deposition: The sequences reported in this paper have been deposited in the GenBank database (accession nos. JQ679485–JQ680031).

[1]C.S. and A.K. contributed equally to this work.

[2]P.Z., D.M., and D.A. contributed equally to this work.

[3]To whom correspondence may be addressed. E-mail: jitender.dubey@ars.usda.gov or sibley@borcim.wus.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1203190109/-/DCSupplemental.

distinction among strains (15). Finally, sequenced-based markers (based on introns in housekeeping genes) have been used to detect polymorphisms and to generate coalescence estimates (6, 9). Although each of these genotyping strategies has advantages, they have largely been used independently, complicating comparison of strains analyzed by different methods. Additionally, the wider application of these typing methods has led to rapid expansion of the number of strains with distinct "genotypes," without improving our understanding of how they are related. Here, we used all three typing methods to provide a comparative analysis of a large number of isolates. We were interested in several goals: (*i*) comprehensively assess strain abundance and geography, (*ii*) compare methods of analysis to enable cross-referencing of methods, (*iii*) cluster genotypes into major groups of related strains, (*iv*) estimate shared ancestry between major groups, and (*v*) identify the prototype strains of each major group for comparative genomic analysis. These analyses provide a framework for considering the global population structure of *T. gondii* and define lineages for future exploration of biological traits and for deep sequencing efforts of representative isolates.

Results

Distribution and Population Structure of Major *T. gondii* Genotypes Based on RFLP Markers. Previous studies have described the isolation and preliminary genotyping of a large number *T. gondii* strains isolated from humans and animals from around the world (Fig. S1). Here, an unprecedented 956 of these isolates were compared using a set of 11 RFLP markers scattered across 8 of 14 chromosomes and plastid genome (Dataset S1). Network analysis revealed two, markedly distinct patterns: some major clusters defined overrepresented, highly clonal genotypes (large circles in Fig. 1); others were unique, or only infrequently sampled (small circles in Fig. 1). The most abundant genotypes correspond to the previously characterized haplogroups, which represent assemblages of closely related strains as defined previously (9). Particularly abundant where haplogroups 2 and 3, which are widely distributed in North America and Europe (9, 10, 16, 17) (Fig. 1). Somewhat surprisingly, these highly clonal groups showed several closely related and highly abundant clusters that were found in both North America and Europe. Because these analyses are based on RFLP markers that do not capture the diversity of surrounding genomic regions, we do not know whether these represent single mutations that have arisen by drift, or whether they represent greater genetic diversity that might have occurred from recombination with a distinct genotype. Other major clusters include haplogroup 12 (6), whereas the highly virulent type I haplogroup (18), was less abundant (Fig. 1). Consistent with previous reports, the clonal lineages (types 1, 2, and 3) predominate in North America and Europe, and wider surveys of strains from Europe have affirmed both the paucity of diversity and the great prevalence of type 2 strains there (16, 17, 19). Parasites that correspond to clonal haplogroups 1, 2, and 3 (by means of RFLP markers) also occur in South America, where more divergent strain types abound (20–22). These clonal isolates may have been recently introduced in the south by, for example, the exchange of agricultural animals or natural migration of birds. However, some or all of these may be misclassified owing to the relatively low level of resolution afforded by RFLP markers initially designed to discriminate among types 1, 2, and 3. Among predominantly South American strains, haplogroups 6 and 9 exemplify clusters of highly similar genotypes, whereas all other isolates comprise smaller clusters representing less frequent genotypes that are related by a dense network (small connected circles in Fig. 1). Although sampling is less complete in Asia, a dominant cluster defined by haplogroup 13 represents strains that appear to be common in China (23, 24). Overall, the highly clonal structure of populations in North America and Europe contrast markedly with much more divergent groups in South America.

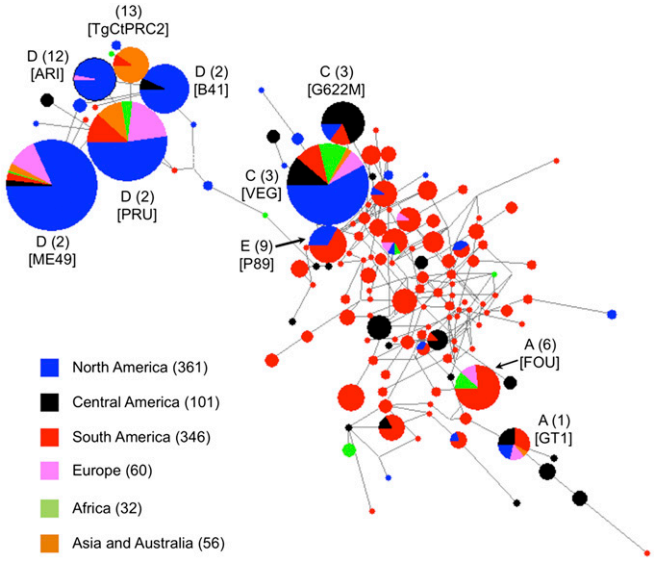

Fig. 1. Geographic distribution of major haplotypes of *T. gondii*. A total of 956 *T. gondii* strains were characterized using 11 multilocus RFLP markers plus 1 marker for the apicoplast (Dataset S1), and a phylogenetic network was constructed using median-joining algorithm implemented in NETWORK 4.1. Samples were collected from North America, Central America, South America, Europe, Africa, Asia, and Australia (numbers of samples are in parentheses). Each circle indicates a unique haplotypes of *T. gondii*. Circle size indicates the number of samples present in each haplotype. Major clades are indicated in uppercase letters, haplogroups indicated in parentheses, and a representative strain is denoted in brackets for each of the major clusters.

Multilocus Analysis of Unique *T. gondii* Genotypes Defined by Network Phylogeny. A total of 138 unique RFLP genotypes were recognized among the 956 strains analyzed above. Representatives of these 138 genotypes were analyzed with three different multilocus sets of markers, including RFLP (25), microsatelites (26), and sequencing of introns from housekeeping genes (9). In total, these markers survey polymorphisms at 30 loci distributed across all 14 chromosomes and the apicoplast (Dataset S1 and S2). Because RFLP and intron-based markers sample single nucleotide changes that arise by mutation, they were combined for analysis. A network derived from the combined polymorphism data from the RFLP and intron-based markers agreed closely with previously defined haplogroups constructed from intron sequencing alone (numbers marked within circles in Fig. 2*A*). However, several groups previously considered distinct were merged: notably haplogroups 4 and 8 (prominent in South America) comprise a single, highly diverse group. Similarly, haplogroups 6 and 14, representing strains found in Africa and South America (10, 17), form part of a major region of the network that also includes the clonal haplogroup 1 (Fig. 2*A*). Haplogroups 2 and 12, related clonal groups common in North America (6), comprise a single major branch. Not anticipated was a cluster of strains related to haplogroups 5, and 10, originally isolated from French Guiana (27, 28), but occupying a distinct position on the network: these strains, which originate from Brazil, have been named haplogroup 15 (Fig. 2*A*). A number of strains were found on long branches that occupy intermediate positions on the network; in some cases, these are represented by individual strains, but others represent strains previously designated as distinct haplogroups (i.e., 7, 11, 13). Although not

MICROBIOLOGY

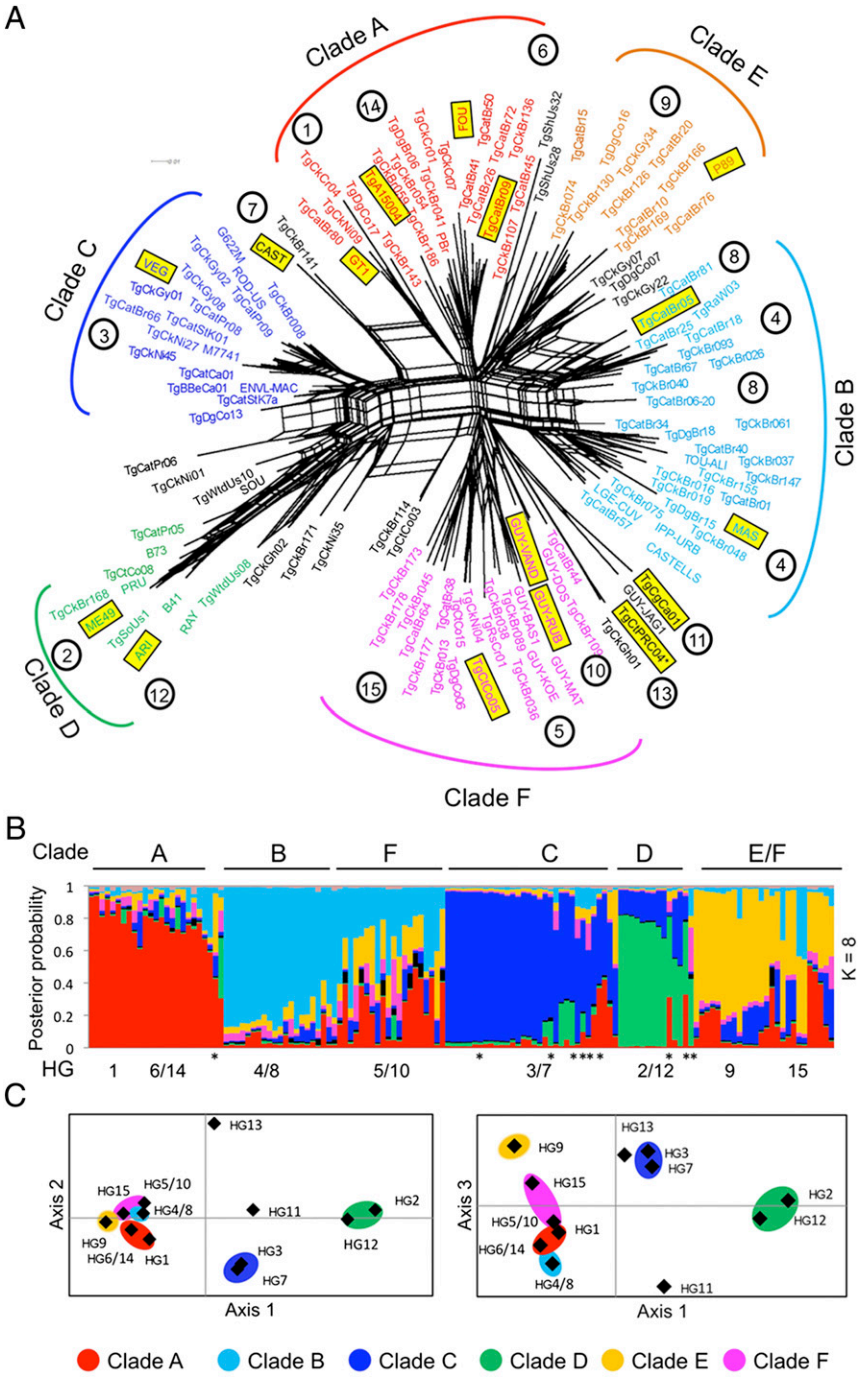

Fig. 2. Population genetic structure of *T. gondii*. (*A*) Neighbor-net analysis was conducted using 11 multilocus RFLP markers plus 1 marker for the apicoplast and four intron sequences from the 138 representative strains representing unique haplotypes. Neighbor-net analysis showed various routes for gene flow between different populations (interconnecting lines between representative strains). Six major clades (A through F) are indicated on the basis of STRUCTURE analysis (Fig. 2*B*). Strains in black lettering do not correspond with major clades. Haplogroups are shown in circled numbers on the basis of previous designations (9, 10). Representative strains for each haplogroup are indicated by yellow boxes. (*B*) Population structure analysis of 138 unique *T. gondii* haplotypes carried out using STRUCTURE. Ancestral population size of 8 ($K = 8$) was chosen as the best fit for the current data (Fig. S2). Population structures of other K values are depicted in Fig. S3. (*Upper*) Clades; (*Lower*) haplogroups (HG). Intermixed haplogroups were separated by / (i.e., 6/14). *, strains that did not correspond closely to their position on the network in Fig. 2*A*. (*C*) Population structure based on PCA analysis. Haplogroups are indicated by numbers. Percentages of variation explained by axes 1, 2, and 3 are 44, 23, and 12, respectively (cumulative ~80%).

commonly encountered, we have retained these haplogroup names to facilitate comparison with prior studies.

Defining Major Clades of Related *T. gondii* Haplogroups. To better define the relationships among haplogroups, we used a Bayesian clustering method called STRUCTURE (29) and principle coordinate analysis (PCA) to group them into clades, which represent groups of related haplogroups, on the basis of the combined RFLP and intron data. For STRUCTURE analysis, we performed 20 simulations of different ancestral population values of K from 1 to 10 and then used the average membership coefficients to generate a plot of their overall relationships. Analysis of combined data from the 138 strains using a previously described method (30) indicated that $K = 8$ ancestral types best explained the current population structure (Fig. S2), although similar profiles were obtained using K values of 5–8 (Fig. 2*B* and Fig. S3). Overall, STRUCTURE organized the 138 unique genotypes into six major clades: haplogroups 1, 6, and 14 define clade A; haplogroups 4 and 8 comprise clade B; haplogroup 3 forms clade C; haplogroups 2 and 12 form clade D; and haplogroup 9 forms clade E (Fig. 2*B*). When these five clades were placed on the network in Fig. 2*A*, they conformed closely to the major branches of the network. In addition, there were several major clusters of strains that were genetically more diverse and not easily classified by STRUCTURE. For example, haplogoups 5 and 10 were found proximate to clade B, on the basis of a shared genetic composition with haplogroups 4 and 8. Somewhat

related genotypes were seen in haplogroup 15, which is found proximate to clade E (Fig. 2*B*). On the basis of their similar mixed genetic composition and proximity in the network, haplogroups 5, 10, and 15 were designated as clade F (Fig. 2 *A* and *B*). Notably, PCA analysis revealed similar relationships for these respective groups (Fig. 2*C*), reinforcing the assignment of specific haplogroups into clades. Overall, the grouping of genotypes by network analysis and STRUCTURE agreed for all but 10 of 138 strains; these exceptions were strains with highly atypical genotypes (indicated in black lettering in Fig. 2*A* and with an asterisk in Fig. 2*B* and Fig. S3). Although these six clades can largely define the major relationships among haplogroups, there is also evidence for shared ancestry between members of clades A and F (Fig. 2).

Detecting Substructure Within Clades. To better define the relationship among haplogroups within clades, we combined the microsatellite polymorphism data with the RFLP and intron polymorphisms and derived separate networks for each clade. For this analysis, we only included strains belonging to haplogroups that define the major clades, excluding strains that lie at the boundaries, occupy rare haplogroups such as 7 and 11, or where classification by network and STRUCURE differed (asterisk in Fig. 2*B*). We also generated estimates of the variance within vs. between groups using F_{ST} as a measure of population subdivision. Assignment of strains to a specific haplogroup was based on reference strains that had been previously genotyped (9, 10) combined with the position of newly sampled strains on the networks (indicated in Dataset S2). Comparison of the network for clade A revealed that haplogroup 1 formed a distinct cluster, separate from strains of haplogroups 6 and 14 (Fig. 3). Each of these haplogroups showed within population heterogeneity (Fig. 4*A*) and separation between these groups was also supported by moderately high F_{ST} values (Fig. 4 *B* and *C*). Clonal type I strains are not abundant and so far have largely been sampled from North America, whereas type 6 and 14 are seen in South America and Africa (10), suggesting the basis for separation is partially geographic. Within the network for clade B, there was a clear bifurcation of strains (Fig. 3), justifying their further subdivision into haplogroups 4 and 8. Analysis of pairwise differences between these groups provided support for population subdivision on the basis of moderately elevated F_{ST} values (Fig. 4). Clade F also showed evidence of subdivision on the basis of network and F_{ST} analysis (Figs. 3 and4). Previously designated haplogroups 5 and 10 were separated in the network (Fig. 3), although less clearly distinct from each other on the basis of F_{ST} values (Fig. 4*C*). In contrast, the newly designated haplogroup 15 formed a separate subgroup within the network (Fig. 3) as supported by elevated F_{ST} values (Fig. 4*C*). Clade D, which is the most divergent from the other clades, showed less evidence of partitioning between haplogroups 2 and 12 (Fig. S4), and low divergence on the basis of F_{ST} (Fig. 4*C*). However, this may result from the choice of markers used here, as previous studies using antigen-encoding genes have documented strong evidence for separation of these groups (6). Clades C and E comprised heterogeneous, but relatively unstructured genotype assemblages (Fig. S4). Because microsatelite markers can evolve at different rates than the other markers, we also analyzed the same set of strains but without inclusion of the microsatellite markers. The separation of clades into distinct haplogroups was also supported by F_{ST} analysis of data that excluded the microsatellites (Fig. S5).

Discussion

We have examined the genetic diversity of *T. gondii* by sampling more than 950 isolates collected from around the world. Our findings reveal that the overall population structure consists of clusters of highly abundant, overrepresented clonal genotypes intermixed with more diverse groups that show greater evidence of outcrossing. The bistability of this pattern is reinforced by strong geographic separation, with clonal isolates being widespread in North America and Europe, and more diverse genotypes found in South America. Of 138 distinct genotypes, most could be grouped into a small number of haplogroups comprising six major clades. Analysis of the relationships among these groups suggests that interbreeding of a small number of founding populations likely gave rise to extant diversity of *T. gondii*, by a process of admixture. Although such patterns are informative on a global level, they should be interpreted with caution as strains that appear genetically similar here may not be genealogically the same due to the limited nature of sampling and the inherent variability of inheritance among progeny of any given genetic cross.

Collectively, this analysis provides a framework for considering genetic diversity and for grouping strains on the basis of shared ancestries. Highly abundant clonal genotypes continue to dominate in North America and Europe, but these results also affirm a growing realization that the extent and structure of diversity is markedly different elsewhere. The contrast is especially evident in South America, where strains show greater evidence of recombination (5).

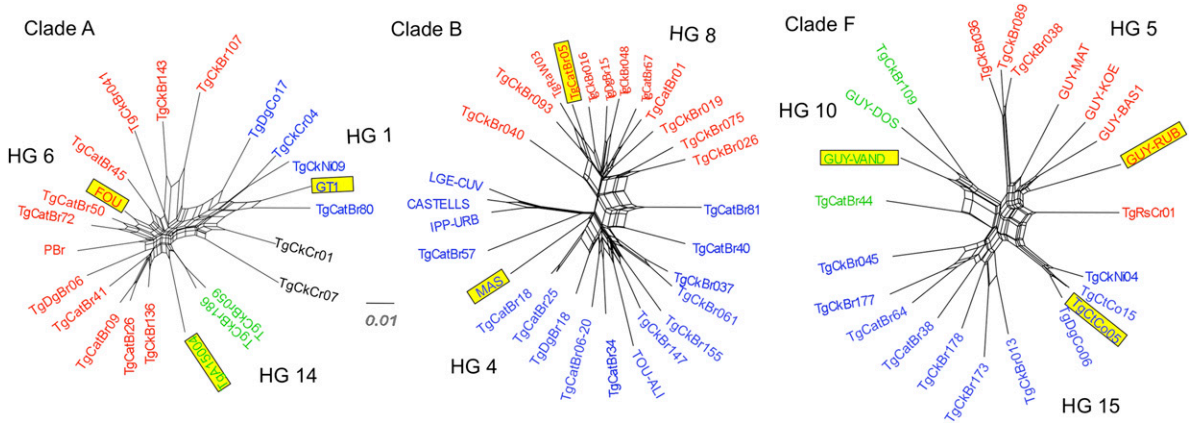

Fig. 3. Neighbor-net analysis of clades using polymorphisms from microsatellite, RFLP, and intron sequences. Haplogroups were analyzed using separate networks on the basis of the clades defined in Fig. 2. Members of each haplogroup were defined by prior designation of reference strains combined with the partitioning of new strains on the network (Dataset S2). Haplogroups designated by different letter coloring; black indicates strains where the clustering here did not coincide with position on Fig. 2*A*. Representative strains for each haplogroup are indicated by yellow boxes.

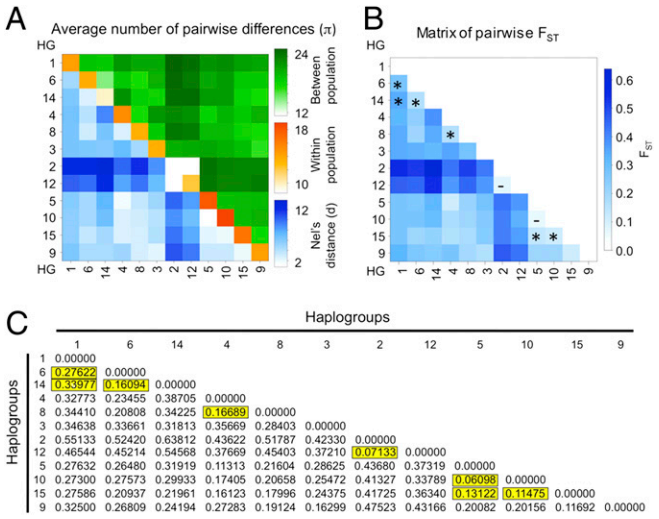

Fig. 4. Graphic representation of population relationships as described by average number of pairwise differences (π) and a matrix of pairwise F_{ST} values. (*A*) Orange-red on diagonal corresponds to pairwise differences (π) within populations, green *Above* diagonal corresponds to between population (*x* and *y* axes), blue *Below* diagonal corresponds to genetic distance (*d*) between populations. (*B*) Population relatedness was determined by F_{ST} between pairs of haplogroups. Pairwise F_{ST} values between haplogroups are depicted. *, moderate value of F_{ST} (i.e., >0.1–0.2), supporting moderate genetic separation; –, F_{ST} value below 0.1, supporting minimal genetic differentiation; HG, haplogroup. (*C*) Table of pairwise F_{ST} values shown in color scheme in *B*. Analyses include polymorphisms defined by microsatellite, RFLP, and intron sequence data.

Ours is a unique study in using all three principal genotyping methods to describe and compare population diversity and structure and represents the broadest population sampling effort to date. Although the identified patterns are likely to change with additional sampling, we anticipate that the current analysis will stand as a useful framework for future studies.

Notably, the genetic diversity of *T. gondii* can be explained as having resulted from admixture among a small number of founders. Previous studies have suggested four to six founding lineages (6, 9, 10); the present work suggests that as many as eight may have contributed. Regardless of the exact number, it is noteworthy that under a range of assumptions, only five major ancestral types, each represented as color pattern in STRUCTURE, explain the vast majority of extant variation. These dominant ancestral types in STRUCTURE parallel the grouping via network analysis, corresponding to five of the major clades (A–E), whereas the sixth (F) is a mixture of other types. Among the six clades, several were shown to comprise single haplotypes such as 3 (clade C) and 9 (clade E), both previously reported to be clonal (9); although here these groups also clearly contain some related yet divergent genotypes that likely reflect recombination and/or mutational drift. Clade D, which contains two related clonal haplogroups, 2 and 12 (6), shows the highest level of divergence from other groups; it is also highly cohesive in STRUCTURE and shows a long-branch profile in network analysis. Clades A, B, and F also show considerable substructure being composed of distinct haplogroups. Clade A represents the group with the widest distribution, being previously described in Africa (haplogroups 6 and 14) (10, 31), South America (haplogroup 6) (9), and North America (haplogroup 1) (32). Shared ancestry is apparent between these haplogroups at all *K* values in STRUCTURE (Fig. S2). This pattern is consistent with the recent suggestion that a type 6 strain may have contributed to the recent origin of the type I lineage in North America (10). Clade B, composed of haplogroups 4 and 8, is almost exclusively confined to South America, where it has previously been associated with ocular disease (33). Finally, clade F is the most diverse genetically, being composed of strains previously isolated from the Amazonian part of French Guiana (haplogroups 5 and 10) (34) along with a unique group of strains (haplogroup 15) primarily originating from domestic animals in Brazil. This group includes chicken isolates from the states of Bahia and Ceara in the east, and cat isolates from Rio de Janeiro and Sao Paulo in the southeast of Brazil, indicating it is widespread.

Marked phenotypic variation occurs among isolates of *T. gondii* in traits governing transmission, virulence, and antigenicity, and such attributes may be shared among members of clonal groups (5). Progress in understanding the mechanistic basis for heritable variation in traits relevant to transmission and pathogenesis requires a robust evolutionary framework for classifying isolates. For example, the acute virulence of haplogroup 1 has been linked to two polymorphic rhoptry kinases (35, 36), and collectively these genes likely explain the majority of this trait for all of the clonal type I strains. Additionally, one of these genes, *ROP18*, also likely contributes to virulence in strains from haplogroups 4, 5, and 10 (20). However, biological phenotypes such as acute virulence in laboratory mice, efficiency of chronic infection, and transmission potential have not been fully evaluated for most of the haplotypes, nor have forward genetic studies been conducted to assess the genetic basis of such traits. One of the advantages of the combined analyses provided here is that it allows for definition of reference strains that represent either diverse genetic groups or highly abundant genotypes. To facilitate future studies on the biology of *T. gondii*, we have designated reference strains for each haplogroup (boxed in Figs. 2*A* and 3 and Fig. S4). These reference strains are currently part of whole genome sequencing effort on the basis of NextGen sequencing and assembly of independent genomes (http://gsc.jcvi.org/projects/gsc/t_gondii/). The genome assemblies and annotations will be deposited in ToxoDB (http://toxodb.org/toxo/) and the strains made available through BEI resources (http://www.beiresources.org/) and at Biological Resource Center Toxoplasma (www.toxocrb.com). Collectively, the results provided here will provide a reference for future comparative genetic analysis of *T. gondii* and for expanding biological and genetic studies on the basis of representative isolates of major lineages.

Methods

Culture of *T. gondii* Strains. *T. gondii* strains were grown in monolayers of human foreskin fibroblast cells and harvested as described previously (9). To prepare templates for PCR, parasites were lysed with 10 μg/mL proteinase K (Sigma) at 55 °C for 2 h and heat inactivated at 95 °C for 15 min (37).

Analysis of DNA Polymorphisms. ***RFLP markers.*** All 956 strains were typed using 11 previously described RFLP markers distributed across 8 of 14 chromosomes plus 1 marker for the apicoplast (25, 38). Strains analyzed here were named using a convention of Tg (for *T. gondii*) followed by a two-letter code for host and a two-letter code for country, followed by an isolate number, or on the basis of previous studies (Dataset S2). DNA sequences were first amplified by multiplex PCR using external primers for all markers followed by nested PCR for each marker separately and analyzed as described previously (25).

Microsatellite markers. Select strains representing the 138 unique genotypes were also typed using 15 microsatellite markers distributed on 10 of 14 chromosomes (Dataset S1), as described previously (26). Samples were analyzed using an automatic sequencer (ABI PRISM 3130xl; Applied Biosystems) and the sizes of the alleles in base pairs estimated using GeneMapper analysis software (version 4.0; Applied Biosystems).

Intron markers. Select strains representing 138 unique genotypes were also typed by sequencing four introns from three different genes (i.e., *UPRT*, *EF*, and *HP*) comprising 1775 bp, by modification of a previously reported method (9) (Dataset S1). PCR-amplified regions were sequenced using BigDye (Applied Biosystems) conducted by Genewiz. Sequences were aligned using Clustal W/X (39) with default settings. Aligned sequences were directly incorporated into molecular evolutionary genetic analysis (MEGA) version 3.1 (40) for identification of single nucleotide polymorphisms (SNPs). Intron

sequences used in the present study have been deposited with GenBank (accession nos. JQ679485–JQ680031).

Network Analysis. Allelic data from multilocus analysis of PCR-RFLP markers were used to generate a network of *T. gondii* strain polymorphisms from 956 samples using the median-joining algorithm (41) (with $\varepsilon = 0$) as implemented in NETWORK 4.1. Multilocus PCR-RFLP typing data were coded for all genetic loci. For a given locus, presence or absence of DNA restriction fragments was coded as either 1 or 0, respectively. SNPs defined by differences in the intron sequences were coded as 1 vs. 0 for all strains. Combined data from PCR-RFLP and intron SNPs were analyzed using SplitsTree v4.4 to compute an unrooted network using the neighbor-net method with 1,000 bootstrap replicates (42). For some networks, microsatellite genotypes were coded as strings of 0 vs. 1 to represent different alleles and added to the analysis of RFLP and intron poylmorphisms.

STRUCTURE Analysis. Polymorphisms from RFLP markers and intron sequence data were concatenated to form a data file. Clustering analysis was carried out using a Bayesian statistical approach implemented in STRUCTURE v2.3.3 (29). Twenty simulation runs were conducted for each of $K = 1$ to $K = 10$ using a length of burn-in of 10^4 and 10^4 replicates of Markov chain Monte Carlo simulation. The simulation was conducted using the linkage model with independent allele frequency for estimating the ancestral populations. The average membership coefficients for the 20 simulation runs of a given K value were generated by CLUMPP v1.1.2 (43) and a graphical presentation of the average membership coefficients for each isolate was generated in Microsoft Excel. An estimate of the true number of populations, K, was calculated using an ad hoc statistic-based approach implemented in software program Structure Harvester v0.6.1, as described previously (30).

PCA. Concatenated SNPs data from RFLP and intron sequences were used for PCA analysis using pairwise population matrix of mean population haploid genetic distance, calculated as described previously (44).

Genetic Distance and F_{ST} Calculation. The average number of pairwise differences (π) was calculated within and between haplogroups and average genetic distance (d), and pairwise F_{ST} values between different haplogroups of *T. gondii* strains were calculated in Arlequin v3.5, on the basis of 10,000 permutations. Graphics were generated automatically by R-lequin using a series of R scripts implemented in Arlequin v3.5.

ACKNOWLEDGMENTS. We thank Jon Boyle, Michael Grigg, David Roos, and Jeroen Saeij for their advice and the Biological Resource Center Toxoplasma network for supplying some strains used here. This work was supported by Grant AI059176 from the National Institutes of Health (to L.D.S.).

1. Dubey JP (2010) *Toxoplasmosis of Animals and Humans* (CRC, Boca Raton), p 313.
2. Levine ND (1988) *The Protozoan Phylum Apicomplexa* (CRC, Boca Raton).
3. Nicolle C, Manceaux LH (1908) On a Leishman body infection (or related organisms) of the gondi. *C R Acad Sci Ill*, 147:763–766. French.
4. Splendore A (1908) A new parasite of rabbits detected in anatomic lesions closely resembling human Kala-azar. *Rev. Soc. Scient. Sao Paulo*, 3:109–112. Italian.
5. 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.
6. 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.
7. 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 USA* 103:10514–10519.
8. Su C, et al. (2003) Recent expansion of Toxoplasma through enhanced oral transmission. *Science* 299:414–416.
9. Khan A, et al. (2007) Recent transcontinental sweep of *Toxoplasma gondii* driven by a single monomorphic chromosome. *Proc Natl Acad Sci USA* 104:14872–14877.
10. 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.
11. Bahia-Oliveira LM, et al. (2003) Highly endemic, waterborne toxoplasmosis in north Rio de Janeiro state, Brazil. *Emerg Infect Dis* 9:55–62.
12. Bowie WR, et al.; The BC Toxoplasma Investigation Team (1997) Outbreak of toxoplasmosis associated with municipal drinking water. *Lancet* 350:173–177.
13. Mead PS, et al. (1999) Food-related illness and death in the United States. *Emerg Infect Dis* 5:607–625.
14. Su C, Zhang X, Dubey JP (2006) Genotyping of *Toxoplasma gondii* by multilocus PCR-RFLP markers: A high resolution and simple method for identification of parasites. *Int J Parasitol* 36:841–848.
15. Ajzenberg D, Bañuls AL, Tibayrenc M, Dardé ML (2002) Microsatellite analysis of *Toxoplasma gondii* shows considerable polymorphism structured into two main clonal groups. *Int J Parasitol* 32:27–38.
16. Ajzenberg D, et al. (2002) Genotype of 86 *Toxoplasma gondii* isolates associated with human congenital toxoplasmosis, and correlation with clinical findings. *J Infect Dis* 186:684–689.
17. 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.
18. Sibley LD, Boothroyd JC (1992) Virulent strains of *Toxoplasma gondii* comprise a single clonal lineage. *Nature* 359:82–85.
19. Dardé ML, Bouteille B, Pestre-Alexandre M (1992) Isoenzyme analysis of 35 *Toxoplasma gondii* isolates and the biological and epidemiological implications. *J Parasitol* 78:786–794.
20. 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.
21. Lehmann T, Marcet PL, Graham DH, Dahl ER, Dubey JP (2006) Globalization and the population structure of *Toxoplasma gondii*. *Proc Natl Acad Sci USA* 103:11423–11428.
22. 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.
23. Dubey JP, et al. (2007) Genetic and biologic characterization of *Toxoplasma gondii* isolates of cats from China. *Vet Parasitol* 145:352–356.
24. Zhou P, et al. (2010) Genetic characterization of *Toxoplasma gondii* isolates from pigs in China. *J Parasitol* 96:1027–1029.
25. Su C, Shwab EK, Zhou P, Zhu XQ, Dubey JP (2010) Moving towards an integrated approach to molecular detection and identification of Toxoplasma gondii. *Parasitology* 137:1–11.
26. Ajzenberg D, Collinet F, Mercier A, Vignoles P, Dardé ML (2010) Genotyping of *Toxoplasma gondii* isolates with 15 microsatellite markers in a single multiplex PCR assay. *J Clin Microbiol* 48:4641–4645.
27. Carme B, et al. (2002) Severe acquired toxoplasmosis in immunocompetent adult patients in French Guiana. *J Clin Microbiol* 40:4037–4044.
28. Dardé ML, Villena I, Pinon JM, Beguinot I (1998) Severe toxoplasmosis caused by a Toxoplasma gondii strain with a new isoenzyme type acquired in French Guyana. *J Clin Microbiol* 36:324.
29. 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.
30. Earl DA, vonHoldt BM (2011) STRUCTURE HARVESTER: A website and program for visualizing STRUCTURE output and implementing the Evanno method. *Conservation Genetics Resources*, 10.1007/s12686-011-9548-7.
31. 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.
32. Howe DK, Sibley LD (1995) *Toxoplasma gondii* comprises three clonal lineages: Correlation of parasite genotype with human disease. *J Infect Dis* 172:1561–1566.
33. Khan A, et al. (2006) Genetic divergence of *Toxoplasma gondii* strains associated with ocular toxoplasmosis, Brazil. *Emerg Infect Dis* 12:942–949.
34. Mercier A, et al. (2011) Human impact on genetic diversity of Toxoplasma gondii: Example of the anthropized environment from French Guiana. *Infect Genet Evol* 11: 1378–1387.
35. Behnke MS, et al. (2011) Virulence differences in *Toxoplasma* mediated by amplification of a family of polymorphic pseudokinases. *Proc Natl Acad Sci USA* 108: 9631–9636.
36. Taylor S, et al. (2006) A secreted serine-threonine kinase determines virulence in the eukaryotic pathogen *Toxoplasma gondii*. *Science* 314:1776–1780.
37. Su C, Howe DK, Dubey JP, Ajioka JW, Sibley LD (2002) Identification of quantitative trait loci controlling acute virulence in *Toxoplasma gondii*. *Proc Natl Acad Sci USA* 99: 10753–10758.
38. Dubey JP, Su C (2009) Population biology of Toxoplasma gondii: What's out and where did they come from. *Mem Inst Oswaldo Cruz* 104:190–195.
39. Higgins DG, Thompson JD, Gibson TJ (1996) Using CLUSTAL for multiple sequence alignments. *Methods Enzymol* 266:383–402.
40. Tamura K, Dudley J, Nei M, Kumar S (2007) MEGA4: Molecular evolutionary genetics analysis (MEGA) software version 4.0. *Mol Biol Evol* 24:1596–1599.
41. Bandelt HJ, Forster P, Röhl A (1999) Median-joining networks for inferring intraspecific phylogenies. *Mol Biol Evol* 16:37–48.
42. Huson DH, Bryant D (2006) Application of phylogenetic networks in evolutionary studies. *Mol Biol Evol* 23:254–267.
43. Jakobsson M, Rosenberg NA (2007) CLUMPP: A cluster matching and permutation program for dealing with label switching and multimodality in analysis of population structure. *Bioinformatics* 23:1801–1806.
44. Huff DR, Peakall R, Smouse PE (1993) RAPD variation within and among populations of outcrossing buffalograss (Buchloë dactyloides (Nutt.) Engelman). *Theor Appl Genet* 96:827–834.

Supporting Information

Su et al. 10.1073/pnas.1203190109

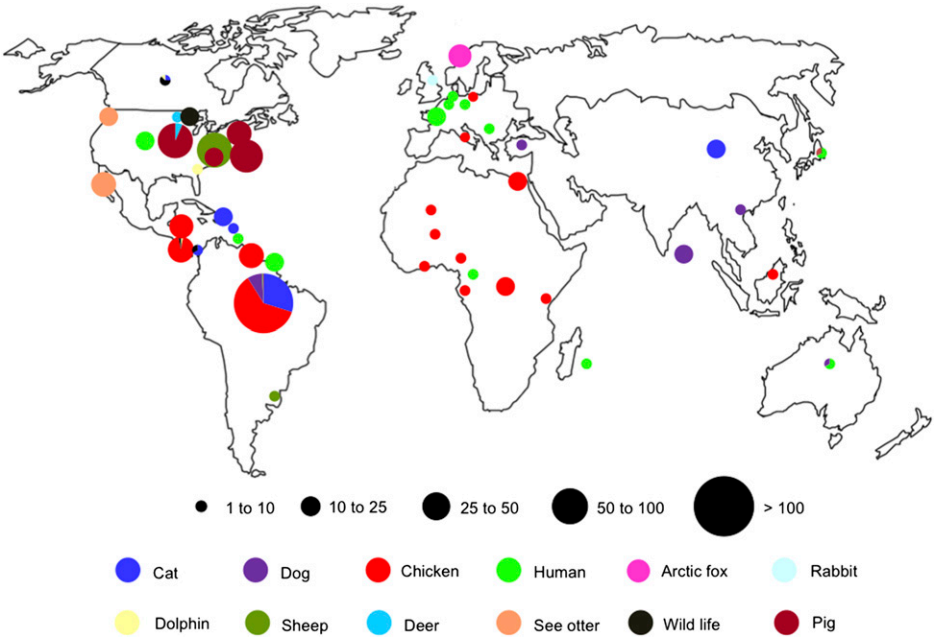

Fig. S1. Worldwide and host distribution of *T. gondii* isolates used in this study. Black circle size indicates the number of isolates. Different color circles indicate the host distribution.

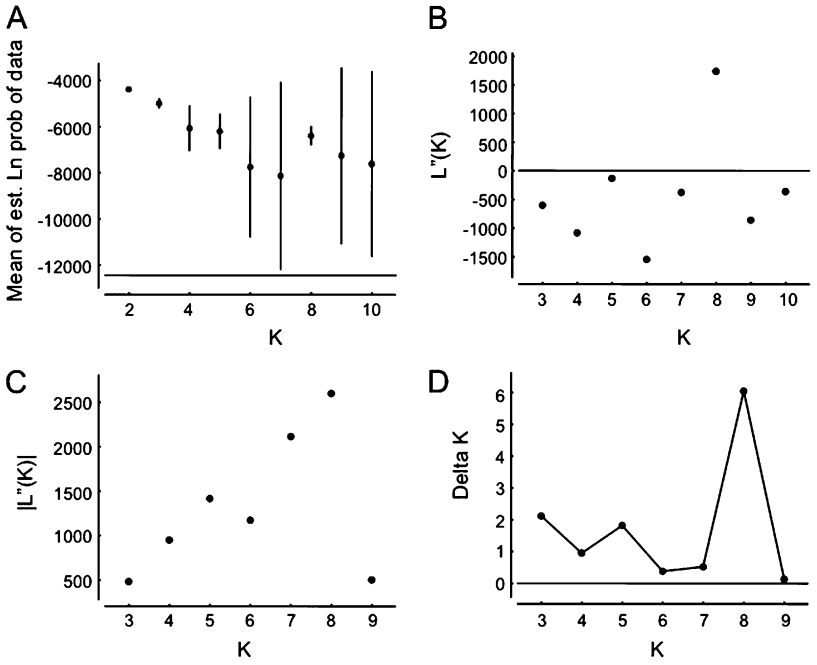

Fig. S2. Estimate of the number of ancestral population (K) and current population structure. (A) Plot of the mean likelihood $L(K)$. (B) Plot of the mean difference between successive likelihood values of K, $L'(K) = L(K) - L(K-1)$. (C) Plot of the mean differences between successive values of $L'(K)$, where $|L''(K)| = |L'(K+1) - L'(K)|$. ($D$) Plot of the delta $K(\Delta K)$, $\Delta K = m|L''(K)|/s[L(K)]$, where m = mean of the absolute values of $L''(K)$, s = SD of $L(K)$.

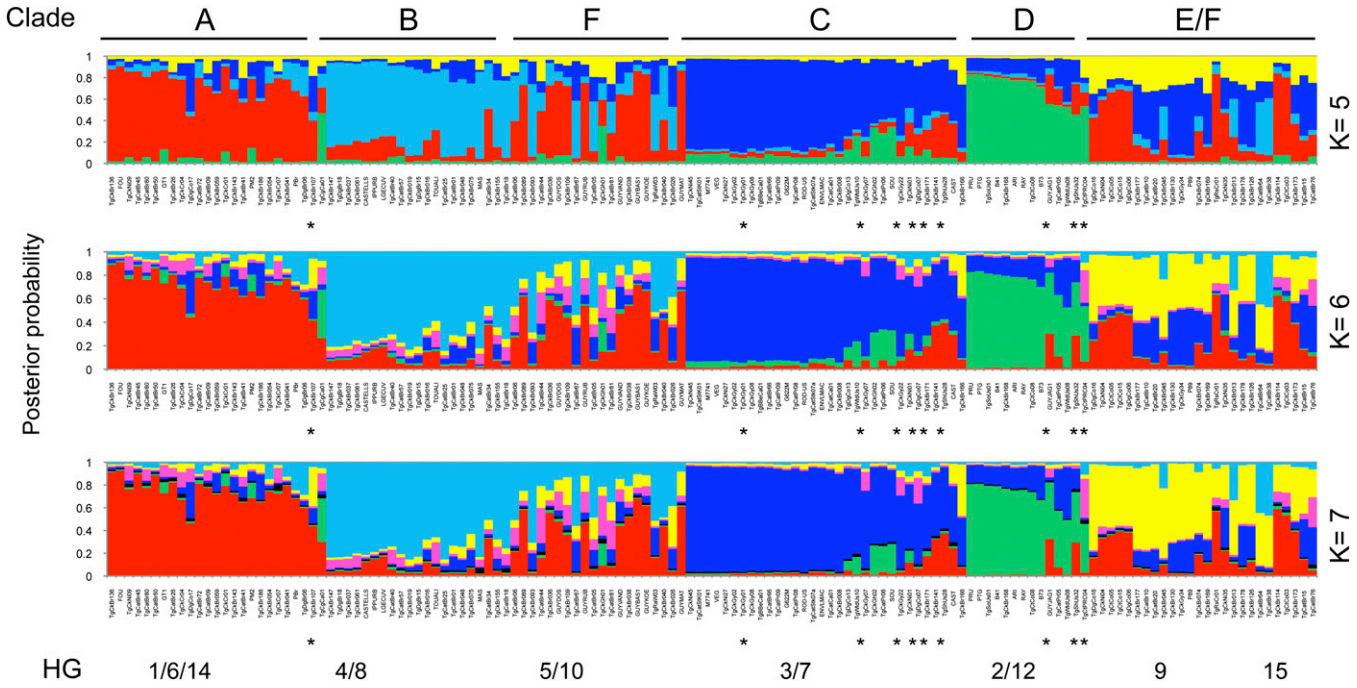

Fig. S3. Population structure based on ancestral population sizes, K = 5–7. Major clades and haplogroups are defined by letters and number, respectively. *, position in STRUCTURE does not correspond closely with the network in Fig. 2*A*.

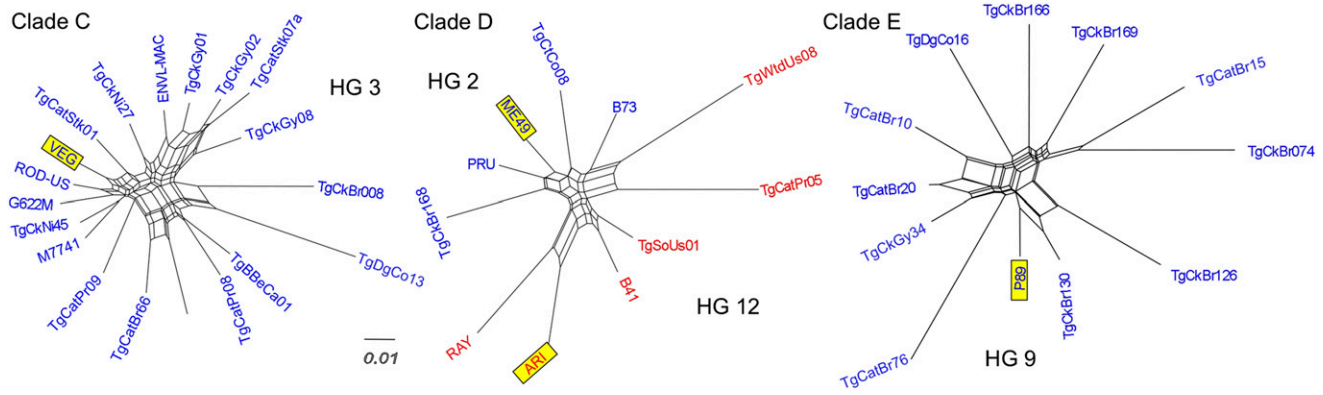

Fig. S4. Neighbor-net analysis of clades using polymorphisms from microsatellite, RFLP, and intron sequences. Members of each haplogroup were defined by prior designation of reference strains combined with the partitioning of new strains on the network (Dataset S2). Haplogroups are designated by different letter coloring. Representative strains for each haplogroup are indicated by yellow boxes.

Haplogroups

Haplogroups

	1	6	14	4	8	3	2	12	5	10	15	9
1	0.00000											
6	0.40759	0.00000										
14	0.52777	0.26247	0.00000									
4	0.51005	0.32421	0.54561	0.00000								
8	0.48120	0.35829	0.48885	0.19400	0.00000							
3	0.49045	0.48607	0.40151	0.52247	0.43105	0.00000						
2	0.73348	0.65613	0.76862	0.64628	0.65688	0.56676	0.00000					
12	0.64671	0.58504	0.67417	0.57697	0.58256	0.52881	0.10659	0.00000				
5	0.46184	0.36594	0.46834	0.21715	0.30934	0.42448	0.64156	0.57738	0.00000			
10	0.41317	0.36126	0.39310	0.27183	0.25176	0.36526	0.57483	0.49210	0.10324	0.00000		
15	0.41624	0.28999	0.33113	0.26162	0.25620	0.40087	0.57847	0.51360	0.19196	0.13711	0.00000	
9	0.47629	0.41946	0.35166	0.44958	0.35043	0.28573	0.65162	0.60300	0.33826	0.30432	0.21228	0.00000

Fig. S5. Table of pairwise F_{ST} values between different haplogroups. Analysis includes polymorphisms defined by RFLP and intron sequence data.

Other Supporting Information Files

Dataset S1 (XLS)
Dataset S2 (XLSX)