Local admixture of amplified and diversified secreted pathogenesis determinants shapes mosaic Toxoplasma gondii genomes

Hernan Lorenzi1,*, Asis Khan2,3,*, Michael S. Behnke2,4,*, Sivaranjani Namasivayam5,6, Lakshmipuram S. Swapna7,8, Michalis Hadjithomas1,6, Svetlana Karamycheva1, Deborah Pinney9, Brian P. Brunk9, James W. Ajioka10, Daniel Ajzenberg11, John C. Boothroyd12, Jon P. Boyle13, Marie L. Dardé11, Maria A. Diaz-Miranda9, Jitender P. Dubey14, Heather M. Fritz15, Solange M. Gennari16, Brian D. Gregory9, Kami Kim17, Jeroen P.J. Saeij18, Chunlei Su19, Michael W. White20, Xing-Quan Zhu21, Daniel K. Howe22, Benjamin M. Rosenthal14, Michael E. Grigg3, John Parkinson7,8, Liang Liu23,24, Jessica C. Kissinger5,6,24, David S. Roos9 & L. David Sibley2

Toxoplasma gondii is among the most prevalent parasites worldwide, infecting many wild and domestic animals and causing zoonotic infections in humans. T. gondii differs substantially in its broad distribution from closely related parasites that typically have narrow, specialized host ranges. To elucidate the genetic basis for these differences, we compared the genomes of 62 globally distributed T. gondii isolates to several closely related coccidian parasites. Our findings reveal that tandem amplification and diversification of secretory pathogenesis determinants is the primary feature that distinguishes the closely related genomes of these biologically diverse parasites. Our further show that the unusual population structure of T. gondii is characterized by clade-specific inheritance of large conserved haploblocks that are significantly enriched in tandemly clustered secretory pathogenesis determinants. The shared inheritance of these conserved haploblocks, which show a different ancestry than the genome as a whole, may thus influence transmission, host range and pathogenicity.

1 Department of Infectious Diseases, The J. Craig Venter Institute, 9704 Medical Center Drive, Rockville, Maryland 20850, USA. 2 Department of Molecular Microbiology, Washington University School of Medicine, 660 S. Euclid Avenue, St Louis, Missouri 63130, USA. 3 Laboratory of Parasitic Diseases, NIAID, National Institutes of Health, Bethesda, Maryland 20892, USA. 4 Pathobiological Sciences, School of Veterinary Medicine, Louisiana State University, Baton Rouge Louisiana 70803, USA. 5 Department of Genetics, University of Georgia, Athens, Georgia 30602, USA. 6 Laboratory for Parasitic Diseases, NIAID, National Institutes of Health, Bethesda, Maryland 20892, USA. 7 Program for Molecular Structure and Function, Hospital for Sick Children, Toronto, Ontario, Canada M5G 1L7. 8 Departments of Biochemistry and Molecular Genetics, University of Toronto, Toronto, Ontario, Canada M5S 1A8. 9 Department of Medicine, University of Pennsylvania, Philadelphia, Pennsylvania 19104, USA. 10 Department of Pathology, University of Cambridge, Cambridge CB2 1QP, UK. 11 Biological Resource Center for Toxoplasma, INSERM, University Limoges, CHU Limoges, UMR_S 1094, Tropical Neuroepidemiology, Institute of Neuroepidemiology and Tropical Neurology, Limoges 87025, France. 12 Department of Microbiology and Immunology, Stanford School of Medicine, Stanford, California 94305, USA. 13 Department of Biological Sciences, Dietrich School of Arts and Sciences, University of Pittsburgh, Pittsburgh, Pennsylvania 15260, USA. 14 Animal Parasitic Diseases Laboratory, Beltsville Agricultural Research Center, Agricultural Research Service, USDA, Beltsville, Maryland 20705, USA. 15 Department of Veterinary Pathology and Microbiology, Washington State University, College of Veterinary Medicine, Pullman, Washington 99164, USA. 16 Department of Preventive Veterinary Medicine and Animal Health, Faculty of Veterinary Medicine, University of São Paulo, São Paulo, SP CEP 05598-270, Brazil. 17 Department of Medicine, Pathology, and Microbiology and Immunology, Albert Einstein College of Medicine, Bronx, New York 10461, USA. 18 Department of Pathology, Microbiology & Immunology, University of California, Davis, California 95616, USA. 19 Department of Microbiology, University of Tennessee, Knoxville, Tennessee 37996, USA. 20 Departments of Molecular Medicine and Global Health, Florida Center for Drug Discovery and Development (CDD), University of South Florida, 3720 Spectrum Boulevard, Suite 304, Tampa, Florida 33612, USA. 21 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, China. 22 Department of Veterinary Science, University of Kentucky, Lexington, Kentucky 40546, USA. 23 Department of Statistics, University of Georgia, Athens Georgia 30602, USA. 24 Institute of Bioinformatics, University of Georgia, Athens, Georgia 30602, USA. * These authors contributed equally to this work. † Present address: Prokaryotic Super Program, DOE Joint Genome Institute, Walnut Creek, California 94598, USA. * Correspondence and requests for materials should be addressed to H.L. (email: hlorenzi@jcvi.org) or to L.D.S. (email: sibley@wustl.edu).
Most of the diversity of eukaryotic life is contained in early branching, unicellular organisms that differ substantially from model organisms such as yeast, flies, worms and mice\(^1\). This diversity is illustrated by the protozoan phylum Apicomplexa, estimated to contain more than 5,000 species\(^2\), most being parasitic on insects and mollusks, while a few cause disease in domestic animals and/or humans\(^3\). Studies of these few disease-causing agents comprise our limited knowledge of this phylum, which demarcate a deep branching phylogeny that has been estimated to span more than \(\sim 400\) my of evolution\(^4\) (Fig. 1a). Over that time frame, it is likely that apicomplexans have adapted to their various vertebrate hosts via multiple independent changes in host range, and yet the molecular mechanisms underlying these adaptations remain largely undefined.

Although most members of this phylum are adapted to a narrow range of hosts, Toxoplasma gondii stands out as a generalist. The genus is characterized by a single species that cause disease in immunocompromised hosts, or as a result of transplacental infection\(^5\). T. gondii is equipped with excellent forward and reverse genetic tools, providing a model for many less-tractable apicomplexan parasites\(^6\). As a highly successful parasite, T. gondii is positioned to inform us about genomic features that are important for efficient transmission and expansion of host range. Here, we sought to exploit this potential by analysing the composition and diversity of the T. gondii genome in comparison to several closely related apicomplexan parasites.

T. gondii belongs to the tissue-cyst forming coccidian parasites, which is distinguished from enteric coccidian parasites by having an alternating two-host (heteroxenous) life cycle (Fig. 1a, Table 1)\(^5\). Most tissue-cyst forming coccidian parasites have obligatory heteroxenous life cycles (that is, Sarcocystis spp. and Hammondia spp.), while others share this mode but have evolved additional strategies for transmission (Table 1)\(^3\). Notably, both T. gondii and Neospora caninum can cause congenital infection, while only T. gondii can be transmitted between intermediate hosts by oral ingestion of infected tissues\(^3\), thus bypassing the sexual phase of the life cycle. These flexible features in the T. gondii life cycle likely aid in transmission through the food chain, thus underlying its broad host range (Table 1). In contrast to our appreciation of differences in life cycle, modes of transmission and host range among these closely related parasites, their molecular bases remain largely unexplored.

In North America and Europe, the population structure of T. gondii is dominated by three prevalent clonal lineages\(^10\), which coexist with much more rare, genetically diverse isolates. A fourth clonal lineage is largely confined to North America, where it is more common in wild animals\(^11\). In contrast, much greater genetic diversity is seen in South America where the population lacks signs of the recent genetic bottleneck and clonal structure seen in the Northern Hemisphere\(^10\). T. gondii utilizes rodents and birds as natural intermediate hosts, and hence it is particularly well adapted for survival in these niches\(^3\). Forward genetic mapping studies have identified several families of secretory proteins in T. gondii that are important for thwarting innate immunity and hence facilitating infection in the mouse\(^12\). Related effectors are conserved in Hammondia hammondi\(^13\), hence the basis for the dramatic differences in biology of these two parasites remains unclear. Nonetheless, one hypothesis advanced by the study of select laboratory strains in the mouse model is that pathogenicity, and perhaps host range, may depend on the repertoire of such secretory pathogenicity determinants, although this has not been tested on a wider level.

Here we tested the generality of this hypothesis through genomic analyses of 62 strains of T. gondii in comparison to several closely related parasites. Our findings reveal that expansion and diversification of secretory pathogenesis determinants (SPDs), which are often tandemly clustered, is a prominent feature of the genomes of T. gondii and related tissue-cyst forming coccidians. Furthermore, patterns of block inheritance, due to recent admixture or selective retention, may underlie specific traits that are shared by related lineages of T. gondii containing similar combinations of SPDs. These features define the population structure of T. gondii and have implications for the evolution of transmission, host range and pathogenicity.

**Results**

**Comparative genomics of tissue-cyst forming coccidians.** We undertook a comparative genomics approach to understand the population diversity of T. gondii and its relationship to closely related tissue-cyst forming coccidian parasites. First, we generated additional genomic DNA sequence coverage (\(~26\times\) coverage) and RNA-seq data (\(>1,000\) × mean coverage of coding sequence) to improve the assembly and annotation for the reference ME49 strain of T. gondii (Table 2). We also generated a whole-genome sequence for H. hammondi (\(~66\times\) coverage; Table 2) and compared these two closely related parasites to the recently completed genomes of Sarcocystis neurona\(^14\) and N. caninum\(^15\) (Table 2), which cause economically important diseases in horses and cattle, respectively (Table 1). Finally, to provide insight into genetic variation of T. gondii we derived whole-genome sequences for 61 additional isolates that were chosen to span presently known global diversity\(^16\) (Supplementary Data 1). Among the total of 62 T. gondii strains, 16 reference strains representing the major haplogroups were sequenced by both 454 (3 and 8 kb paired-end libraries) and Illumina (300 bp paired-end libraries) technologies and the resulting reads were assembled and annotated separately (\(~47\times\) average sequence coverage) (Supplementary Table 1). The remaining strains were sequenced using Illumina only (\(~42\times\) average sequence coverage) and were aligned to the reference strain ME49 (Supplementary Data 1). Below, we present the comparative analyses of these genomes focusing on three broad themes: (1) comparison of T. gondii to the most closely related tissue-cyst forming coccidian parasites, (2) analysis of the core genome of T. gondii and how it has diversified and (3) examination of how the global population structure of T. gondii has been shaped by local genomic admixture.

We compared the whole-genome sequences from four related tissue-cyst forming coccidian parasites spanning a range of biological hosts and life-cycle strategies (Tables 1 and 2, Fig. 1). Three of the four organisms have a similar total genome size of 62–65 Mb (N. caninum, T. gondii and H. hammondi), while the S. neurona genome is somewhat larger due to expanded repeats and much larger introns (Table 2, Supplementary Table 1)\(^14\). All four genomes have roughly similar GC compositions and are predicted to encode from 7,000 to slightly more than 8,000 genes located on 14 chromosomes, as verified in T. gondii\(^17\) (Table 2). Similar to other genome sequencing projects, 42–56% of the predicted CDSs (coding DNA sequences) encode genes with a putative functional domain annotation, while 44–58% are hypothetical unknowns. To identify conserved features, we compared the four different genomes to the enteric coccidian Eimeria tenella\(^18\) using OrthoMCL to cluster genes into putative orthogroups\(^19\). Not surprisingly, more closely related taxa showed a higher degree of shared OrthoMCL clusters (Fig. 1b, Supplementary Data 2). Orthogroups classified by Pfam\(^20\) domains and grouped into the top 20 Gene Ontology
Figure 1 | Comparative genomics of tissue-cyst forming coccidian parasites. (a) Phylogenetic tree of selected apicomplexans based on a conserved DEAD box helicase protein (TGM49_249810, OrthoMCL OG5126701). Neighbour-joining tree with bootstrap values indicated. Distance equals 0.1 amino acid substitutions/site. Taxa names from http://tolweb.org/Apicomplexa. (b) Summary of conserved orthologues based on OrthoMCL analysis. Et, E. tenella; Hh, H. hammondi; Nc, N. caninum; Sn, S. neurona; Tg, T. gondii. (c) Abundance of Pfam domains in coccidian parasites. For each species, the incidence of Pfam domains per protein was determined. The top 20 Pfam domains in T. gondii ME49 are shown along with the number of proteins containing these domains in each of the species. The cells in the table are colour coded based on the rank of Pfam domains in terms of their abundance. The domains of interest are indicated by an asterisk. Taxa as in a. (d) Circos plot illustrating levels of synteny among the coccidian parasites. The large outer circle represents the annotated chromosomes or scaffolds of each coccidian species. For T. gondii, H. hammondi and N. caninum all assembled chromosomes (n=14) are plotted. For S. neurona and E. tenella, the largest 14 scaffolds are plotted. Each chromosome/scaffold is labelled with the genus-species abbreviation followed by the chromosome/scaffold number. Tick marks on the chromosome/scaffold represent 1 Mb. The coloured bands and lines linking chromosome/scaffold pairs represent syntenic blocks (minimum of three genes) shared by the chromosomes that are connected. The syntenic links are drawn with T. gondii, H. hammondi and N. caninum as the reference, in that order. Syntenic blocks were generated using genes present in orthologue clusters, where the cluster contained at least one gene from each species.

Table 1 | Summary of the life cycle, host range and pathogenicity of tissue-cyst forming coccidian parasites.

| Organism          | Definitive host | Alternative host(s) | Transmission                        | Animal disease                                  | Human disease       |
|-------------------|-----------------|---------------------|------------------------------------|------------------------------------------------|--------------------|
| Sarcocystis neurona | Opossum         | Skunk, raccoon and other small mammals | Obligatory sexual-asexual cycle | Myelencephalitis in horses and marine mammals | None               |
| Neospora caninum   | Canine          | Bovine              | Vertical asexual or sexual-asexual cycle | Congenital, abortion in cattle | None               |
| Hammondia hammondi | Feline          | Rodents             | Obligatory sexual-asexual cycle    | None                                          | None               |
| Toxoplasma gondii  | Feline          | Warm-blooded vertebrates | Vertical asexual, sexual-asexual cycle or direct asexual transmission | Congenital, abortion in sheep | Opportunistic pathogen in humans; CNS, ocular, congenital |

(30) terms (http://geneontology.org/) revealed that all five species share similar orthologous groups for many key biological functions, and that tissue-cyst coccidians are enriched in processes involved in protein modification (Supplementary Fig. 1).

The most abundant protein domains in tissue-cyst forming coccidian parasites include serine/threonine (S/T) kinases, RNA-binding proteins, PP2C-type S/T phosphatases and calcium-binding motifs (EF-hands) (Fig. 1c). There is a precedent for the importance of S/T kinases11, such as the expanded polymorphic family of rhoptry (ROP) kinase virulence determinants in T. gondii12 and of calcium-binding motifs, including within a family of calcium-dependent protein kinases13. However, the abundance of RNA-binding proteins (RMR and
RMA motifs was unexpected, as these have been largely unexplored in *T. gondii* and closely related parasites. In addition, plant-like AP2 transcription factors are abundant with these being major transcription factors in apicomplexans. Paralogues of AP2-like transcription factors are also prevalent in *H. hammondi*, as noted previously, and a related family is found in other coccidian parasites, although this result is likely due to their divergence from canonical SRS domains, as similar families of 6-Cys rich proteins occur in other coccidian parasites and a related family is found in *Plasmodium*. SRS proteins and related 6-Cys proteins share a common extracellular structural domain, which are amplified and highly divergent among tissue-cyst forming coccidian parasites. These appear less abundant in *S. neurona* and absent in *E. tenella*, forming coccidian parasites. These appear less abundant in *S. neurona* and absent in *E. tenella*, forming coccidian parasites. From the predicted proteomes, we also reconstructed common metabolic pathways, which were highly conserved across *T. gondii*, *H. hammondi* and *N. caninum*, as noted previously. Expanding this analysis to include the 16 reference strains of *T. gondii* identified paralogues for certain functions, for example, in the pyrimidine and purine metabolic pathways and fatty-acid biosynthesis. Most enzymes involved in energy metabolism were well conserved with few paralogues and/or non-synonymous polymorphisms. Previous studies have established the ability of these pathways to mediate strain-specific growth differences. It is therefore interesting to note that several enzymes with the capacity to modulate flux within these pathways were associated with paralogous expansions and/or significant numbers of non-synonymous polymorphisms.

Finally, we compared the position of genes across the chromosomes to establish the extent of synteny. There was a high degree of conservation of chromosomal position of orthologous genes between *T. gondii* and *H. hammondi*, and this only slightly decreased when they were compared with *N. caninum*, as previously reported for comparisons of *T. gondii* and *N. caninum* (Fig. 1a). Analysis of the more complete *H. hammondi* genome provided here revealed that it shares 29 long syntenic blocks with *T. gondii* harbouring >80% of its genes, with only a few blocks rearranged, most notably a ~1 Mb reciprocal translocation between chromosomes Ia and IX. In contrast, synteny broke down substantially when these three organisms were compared with *S. neurona* and was completely absent when compared with *E. tenella*, as has been described previously for a pairwise comparison of *T. gondii* and *E. tenella*. The loss of synteny since the divergence of enteric from tissue-cyst forming coccidians stands in stark contrast to the conservation of synteny in the kinetoplastidae, fungi and chordates, all groups with greater evolutionary divergence times relative to the coccidians.

### Expanded SPDs in *T. gondii*

To highlight key features of the *T. gondii* genome we depicted the coding capacity of the reference ME49 strain as a Circos plot, where the outermost circle indicates the genes encoded by each of the 14 chromosomes (Fig. 2a, Supplementary Fig. 4). By comparing the average sequence read depth across the genome, we identified chromosomal genes with copy-number variation (CNV) (Fig. 2a, second innermost circle). Expanding this analysis to all 62 strains revealed 14 genes that have evidence of CNV in all strains, and 39 genes with CNV in 90% of the strains (Supplementary Fig. 5, Supplementary Data 4). Examination of patterns of CNV also revealed several examples of large segmental duplications or aneuploidy in specific strains (Supplementary Fig. 6), similar to reports from previous genetic crosses. These regions were genetically homogeneous, suggesting they arose by duplication events and are not hybrids created by unequal crossing over at meiosis. The significance of these diploid regions is uncertain, although recent studies in yeast indicate that aneuploidy can accelerate evolutionary adaptation.

Recent comparisons of the draft genomes of *T. gondii* and *H. hammondi* and the published genome of *N. caninum*, highlighted the expansion of gene families that differ between these otherwise closely related species. Using the newly annotated assembly of the ME49 genome obtained here, and data from 61 additional genomes of *T. gondii*, we expanded these analyses to examine the distribution of amplified genes.
evident either as CNV or tandem arrays in the assemblies. These amplified genes were plotted as concentric coloured circles corresponding to the protein families they belong to and using symbols proportional to their total copy number (Fig. 2a). Many of these amplified genes encode secretory or surface proteins that have been previously implicated in host pathogenesis, referred to
here as SPDs. These SPDs include genes encoding secretory proteins found in micronemes (MICs), dense granules (GRA), ROPs, as well as the SRS super family (Fig. 2a, Supplementary Fig. 4). Members of these protein families are known to mediate host cell attachment (MICs), modification of host immunity (GRA and ROP proteins) or adherence and immune evasion (SRS). Within the ME49 reference genome, we detected a total of 57 gene loci with CNV, which contain 176 gene copies, and 95 loci in tandem arrays, which contain 264 gene copies (Supplementary Data 4). Both CNV and tandemly duplicated genes were enriched in SPDs, in particular in genes encoding SAG/SRS, ROP and MICs (Fig. 2b,c, Supplementary Data 4), a pattern also noted previously. For example, SPDs comprised 52 (35%) of the 152 expanded loci and 196 (45%) of the 440 expanded gene copies, despite making up only 375 (4.5%) of the 8,311 total genes in the ME49 genome. Many of the SPDs also show evidence of positive selection, evident in elevated frequencies of non-synonymous (dNS) versus synonymous (dS) mutations (Fig. 3a). Among these, the GRA, ROP and SAG genes show some of the highest levels of dNS/dS, while metabolic enzymes typically show selection for conservation, as seen by low levels of dNS/dS (Fig. 3a, Supplementary Data 5).

We expanded the analysis of SPDs to examine their diversity among a set of reference genomes representing the 16 major haplogroups (Fig. 3b). OrthoMCL clustering of the SPD families revealed that while most members were represented in all 16 haplogroups, differences in representation and copy number were most evident in the SRS and ROP families (Fig. 3b). Collectively,

**Figure 3 | Analysis of expanded or positively selected gene families in T. gondii.** (a) Analysis of positive selective pressure among T. gondii gene families based on dNS/dS. Red values indicate categories with significantly elevated dNS/dS ratios, indicating positive selection for diversification, while blue values indicate categories with significantly reduced values, indicative of selection for conservation. Analysis based on 16 reference T. gondii genomes, although similar patterns were also seen for all 62 strains. Horizontal black line dNS/dS median; yellow square, dNS/dS mean; mean values significantly different than the category ‘Other’ (P ≤ 0.05, Mann–Whitney–Wilcoxon test) are denoted with ‘*’. (b) Heatmap of the abundance of SPDs in conserved orthogroups. The number of genes in orthogroups containing SPD genes was determined across the 16 T. gondii reference genomes, H. hammondi and N. canum, and plotted as a heatmap. Each row is an orthogroup, and the colour of each cell represents the number of genes in that orthogroup, range 0–19. (c) Schematic representation of domain structure of T. gondii protein families A to E. Family names and average protein lengths per family are indicated on the right. HVR, hypervariable region; KRDEGST-rich domain, protein domain rich in polar amino acids Lysine (K), Arginine (R), Aspartate (D), Glutamate (E), Serine (S), Threonine (T) and Glycine (G); SP, signal peptide; TMM, transmembrane domain; TR, tandem-repeat domain; VR, protein domain of variable length and sequence. Sequence motifs for each tandem-repeat unit are depicted in Logo format. Extra (ED) and intracellular (ID) protein domains, as predicted by Phobius, are indicated below protein schemes. (d) Life-cycle expression of TgFAM genes. Heatmap of log2-normalized microarray expression values for TgFAM genes across the T. gondii life cycle. Background values for the array are near log2 of 5. Samples for unsporulated oocysts (Day 0 Oocyst), sporulated oocysts (Day 10 Oocyst), intermediate host tachyzoites (Day 2 Tachy), intermediate host dormant bradyzoite cysts (Day 8 Brady), and enteric stage merozoites (Day 5 Mero) are plotted.
these analyses reveal that the major difference between *T. gondii* strains is the diversification of SPD family members.

Comparison of orthologues for GRA, ROP and SRS genes between *T. gondii*, *H. hammondii* and *N. caninum* revealed substantial differences in clustering by OrthoMCL, suggesting that the divergence among these genes may underlie biological differences between these species (Fig. 3b, Supplementary Fig. 7). In contrast, MIC genes were highly conserved, suggesting these organisms use a similar repertoire of host receptors (Fig. 3b). Comparison of OrthoMCL groupings also identified a number of putative species-specific genes unique to *N. caninum*, *H. hammondii* or *T. gondii* (Supplementary Fig. 8a, Supplementary Data 6). Further analysis indicated that a subset of the putative species-specific genes represent distant orthologues that are classified as separate groups by OrthoMCL (Supplementary Fig. 8b, Supplementary Data 6). Notably, this distant orthologue category is greater when comparing *N. caninum* to either *T. gondii* or *H. hammondii*, versus the pairwise comparison between the later two species. Among these distant related orthologues, a number encode *TgFAM* or SRS genes, consistent with the idea that they influence important aspects of the biology (Supplementary Data 6). In contrast, a large number of the genes that differ between *T. gondii* and *H. hammondii* show evidence of alternative gene models, including early truncations, premature stop codons and frame shifts (Supplementary Fig. 8b, Supplementary Data 6). In addition, a smaller number of genes were present only in one species and are predicted to be unique, the majority of which were annotated as hypothetical unknowns. Analysis of alternative allele frequencies, RNA-seq data, and sequencing depth coverage failed to find evidence that these predicted differences are due to sequencing or assembly errors and instead suggest that many are genuine (Supplementary Fig. 8c–f). Consequently, the putative unique gene list provided in Supplementary Data 6 provides a tentative starting point to identify genes that may mediate important biological differences between these closely related species.

In addition to the previously recognized SPDs, we identified families of genes that are uniquely enriched in the *T. gondii* genome, referred to here as *TgFAM* genes (Fig. 3c, Supplementary Data 7), including one previously referred to as *Toxoplasma*-specific family (*TSF*)48, which corresponds to *TgFAMC* here. Our analysis of multiple *T. gondii* genomes reveals a much broader set of *Toxoplasma*-specific families (*TgFAMs*) (Supplementary Data 7), five of which we have specifically highlighted for their unique domain structures (Fig. 3c). Several *TgFAMs* are expanded and show evidence of CNV and/or tandem duplication, while others are located at the ends of chromosomes (Fig. 2), as previously noted for the *TSF* family34. This pattern of telomeric clustering has also previously been associated with antigenic variant surface adhesins in *Plasmodium*35. Although *T. gondii* is not known to undergo antigenic variation, the variable domains of the *TgFAMs* may represent adaptations to enhance host cell recognition and/or escape immune detection. We have specifically highlighted five of the *TgFAMs* here based on the fact that they contain conserved signal peptides as well as domain architectures that suggest they may encode surface proteins with extracellular domains that contain conserved protein motifs (Fig. 3c). *TgFAM* genes are expanded in *T. gondii*, although they are less common in *H. hammondii* and *N. caninum* (Fig. 3b, Supplementary Data 7). In particular, *H. hammondii* and *N. caninum* contain far fewer members of *TgFAM* and *TgFAMB*, and *TgFAMC* appears to be largely absent in *N. caninum* (Fig. 3b, Supplementary Fig. 7). Notably, many of the *TgFAM* genes highlighted here are expressed during sexual development in the cat gut or in oocysts that are shed into the environment following the sexual phase (Fig. 3d)39, suggesting they may play roles during transmission. In addition to the *TgFAM* genes highlighted here, there are a number of other gene families containing parasite-specific motifs that are expanded in *T. gondii*, and which may contribute to important biological traits not yet identified (Supplementary Data 7).

**Co-inheritance of haploblocks shape population structure.**

Previous studies have reported the influence of recombination on the global population structure of *T. gondii*, which shows marked geographic segregation of major haplogroups16,27,28, although the factors shaping these patterns remain unresolved. To examine the population structure based on genome-wide polymorphism data, we analysed single nucleotide polymorphisms (SNPs) that were defined by comparison of 61 *T. gondii* strains to the reference ME49 genome and filtered this set to include positions where reliable data were available for all strains (a total of 802,764 positions in each genome (Supplementary Data 8)). Generation of a neighbour network39 for these data revealed that the 62 strains group closely with haplogroups and major clades that were previously defined by lower resolution genotyping (Fig. 4a). Importantly, similar groupings were defined using admixture40 (Supplementary Fig. 9) and principal components analysis (Supplementary Fig. 10). Collectively, these findings support a population structure consisting of a small number of clades that show strong geographic segregation, as described previously16,27.

Although the neighbour network permits visualization of gene flow along several pathways, it does not fully capture the extent or pattern of local genomic admixture among any given pair of strains. To illustrate this more directly, we generated pairwise SNP diversity plots for three of the haplogroups contained in clade D, comparing them to the reference strain ME49 (Fig. 4b). Strains like ARI (haplogroup 12), a sister group of type 2 that is also found in North America, contain large haploblocks that are similar to ME49 (~60%), interspersed with regions that are divergent (Fig. 4b), consistent with previous findings that these two groups are closely related11. In contrast, *TgCtPRC2* (haplogroup 13), which is a common clonal genotype in China41, shares fewer regions with ME49 (~40%) and COUG (haplogroup 11), which represents a rare North American lineage found in wild animals, showed almost no conserved regions with ME49 (~1%) (Fig. 4b). Thus although members of a common clade contain distinct genomic patterns that have arisen by different evolutionary paths, it is striking that many share large conserved haploblocks across their genomes.

To better represent the shared ancestry across strains, we analysed local inheritance patterns using chromosome painting to reveal patterns of local admixture. When strains were aligned by clade, the presence of shared haploblocks across members was evident by common colour patterns (Fig. 4c, Supplementary Fig. 11). These shared regions represent chromosomal haploblocks that show a high degree of shared ancestry, in some cases eroding the boundaries of the clade structure. Noteworthy, this analysis also revealed patterns of local admixture that suggest the occurrence of genetic crosses among strains of different clades, likely favoured by their geographic proximity (Fig. 4c, Supplementary Fig. 11). Similar shared chromosomal haploblocks are also seen in average pairwise plots for SNPs among members of individual clades (Supplementary Fig. 12). Among the most strongly conserved haploblocks is chromosome 1a, which is shared across nearly all clades, with the exception of clades E and F (Fig. 4c,d, Supplementary Fig. 13). The basis for the widespread conservation of chromosome 1a (ref. 37) is uncertain, but recent studies suggest that it may be due to the enhanced transmission in domestic cats42.

The analyses presented above suggest that common inheritance of large haploblocks is the major factor in determining the phylogenetic grouping of *T. gondii* strains. To test this model, we
performed two types of analysis to cluster strains using the SNP data. First, we analysed SNPs using the linkage model of ChromoPainter in FineStructure\(^5\) to generate a clustering hierarchy. This model, which combines information across linked markers in a co-ancestry matrix, recreated the clades seen in the previous analysis with several minor exceptions (Supplementary Fig. 14). Separately, we analysed the SNPs using a rolling window method to define an overall similarity index based on how many regions were co-inherited between all strains independently grouped the strains by clade.

Second, we performed two types of analysis to cluster strains using the SNP data. First, we analysed SNPs using the linkage model of ChromoPainter in FineStructure\(^5\) to generate a clustering hierarchy. This model, which combines information across linked markers in a co-ancestry matrix, recreated the clades seen in the previous analysis with several minor exceptions (Supplementary Fig. 14). Separately, we analysed the SNPs using a rolling window method to define an overall similarity index based on how many regions were co-inherited between all strains independently grouped the strains by clade.

where large chromosomal haploblocks have been inherited in common by members of individual clades. Although recent admixture had previously been suggested by analysing individual regions separately\(^3\), the present genome-wide analysis of SNPs establishes that this pattern is a defining feature of the population structure of *T. gondii*.

To further examine the pattern of long-haploblock inheritance, we compared the ancestry of regions that were conserved by members of the same clade, it emerged that discrete regions of the genome show very low SNP diversity, while others are highly variable (Fig. 5a). Regions of low average pairwise SNP diversity were observed in all clades, but differed in their frequency and scale bar, total number of regions separately. The present genome-wide analysis of SNPs admixture had previously been suggested by analysing individual regions separately\(^3\), the present genome-wide analysis of SNPs establishes that this pattern is a defining feature of the population structure of *T. gondii*.

To further examine the pattern of long-haploblock inheritance, we compared the ancestry of regions that were conserved by members of the same clade, it emerged that discrete regions of the genome show very low SNP diversity, while others are highly variable (Fig. 5a). Regions of low average pairwise SNP diversity were observed in all clades, but differed in their frequency and scale bar, total number of regions separately. The present genome-wide analysis of SNPs admixture had previously been suggested by analysing individual regions separately\(^3\), the present genome-wide analysis of SNPs establishes that this pattern is a defining feature of the population structure of *T. gondii*.

To further examine the pattern of long-haploblock inheritance, we compared the ancestry of regions that were conserved by members of the same clade, it emerged that discrete regions of the genome show very low SNP diversity, while others are highly variable (Fig. 5a). Regions of low average pairwise SNP diversity were observed in all clades, but differed in their frequency and scale bar, total number of regions separately. The present genome-wide analysis of SNPs admixture had previously been suggested by analysing individual regions separately\(^3\), the present genome-wide analysis of SNPs establishes that this pattern is a defining feature of the population structure of *T. gondii*.

To further examine the pattern of long-haploblock inheritance, we compared the ancestry of regions that were conserved by members of the same clade, it emerged that discrete regions of the genome show very low SNP diversity, while others are highly variable (Fig. 5a). Regions of low average pairwise SNP diversity were observed in all clades, but differed in their frequency and scale bar, total number of regions separately. The present genome-wide analysis of SNPs admixture had previously been suggested by analysing individual regions separately\(^3\), the present genome-wide analysis of SNPs establishes that this pattern is a defining feature of the population structure of *T. gondii*.

To further examine the pattern of long-haploblock inheritance, we compared the ancestry of regions that were conserved by members of the same clade, it emerged that discrete regions of the genome show very low SNP diversity, while others are highly variable (Fig. 5a). Regions of low average pairwise SNP diversity were observed in all clades, but differed in their frequency and scale bar, total number of regions separately. The present genome-wide analysis of SNPs admixture had previously been suggested by analysing individual regions separately\(^3\), the present genome-wide analysis of SNPs establishes that this pattern is a defining feature of the population structure of *T. gondii*.

To further examine the pattern of long-haploblock inheritance, we compared the ancestry of regions that were conserved by members of the same clade, it emerged that discrete regions of the genome show very low SNP diversity, while others are highly variable (Fig. 5a). Regions of low average pairwise SNP diversity were observed in all clades, but differed in their frequency and scale bar, total number of regions separately. The present genome-wide analysis of SNPs admixture had previously been suggested by analysing individual regions separately\(^3\), the present genome-wide analysis of SNPs establishes that this pattern is a defining feature of the population structure of *T. gondii*.

To further examine the pattern of long-haploblock inheritance, we compared the ancestry of regions that were conserved by members of the same clade, it emerged that discrete regions of the genome show very low SNP diversity, while others are highly variable (Fig. 5a). Regions of low average pairwise SNP diversity were observed in all clades, but differed in their frequency and scale bar, total number of regions separately. The present genome-wide analysis of SNPs admixture had previously been suggested by analysing individual regions separately\(^3\), the present genome-wide analysis of SNPs establishes that this pattern is a defining feature of the population structure of *T. gondii*.

To further examine the pattern of long-haploblock inheritance, we compared the ancestry of regions that were conserved by members of the same clade, it emerged that discrete regions of the genome show very low SNP diversity, while others are highly variable (Fig. 5a). Regions of low average pairwise SNP diversity were observed in all clades, but differed in their frequency and scale bar, total number of regions separately. The present genome-wide analysis of SNPs admixture had previously been suggested by analysing individual regions separately\(^3\), the present genome-wide analysis of SNPs establishes that this pattern is a defining feature of the population structure of *T. gondii*.

To further examine the pattern of long-haploblock inheritance, we compared the ancestry of regions that were conserved by members of the same clade, it emerged that discrete regions of the genome show very low SNP diversity, while others are highly variable (Fig. 5a). Regions of low average pairwise SNP diversity were observed in all clades, but differed in their frequency and scale bar, total number of regions separately. The present genome-wide analysis of SNPs admixture had previously been suggested by analysing individual regions separately\(^3\), the present genome-wide analysis of SNPs establishes that this pattern is a defining feature of the population structure of *T. gondii*.

To further examine the pattern of long-haploblock inheritance, we compared the ancestry of regions that were conserved by members of the same clade, it emerged that discrete regions of the genome show very low SNP diversity, while others are highly variable (Fig. 5a). Regions of low average pairwise SNP diversity were observed in all clades, but differed in their frequency and scale bar, total number of regions separately. The present genome-wide analysis of SNPs admixture had previously been suggested by analysing individual regions separately\(^3\), the present genome-wide analysis of SNPs establishes that this pattern is a defining feature of the population structure of *T. gondii*.

To further examine the pattern of long-haploblock inheritance, we compared the ancestry of regions that were conserved by members of the same clade, it emerged that discrete regions of the genome show very low SNP diversity, while others are highly variable (Fig. 5a). Regions of low average pairwise SNP diversity were observed in all clades, but differed in their frequency and scale bar, total number of regions separately. The present genome-wide analysis of SNPs admixture had previously been suggested by analysing individual regions separately\(^3\), the present genome-wide analysis of SNPs establishes that this pattern is a defining feature of the population structure of *T. gondii*.
To determine the influence of these shared haploblocks on the content of genes found within specific clades, we plotted the distribution of SPDs found within conserved regions shared by members of specific clades (Fig. 5c, Supplementary Data 11). The distribution of SPDs revealed that a number of known pathogenicity determinants were common to conserved regions in specific clades (Fig. 5c). We tested whether these patterns were random, or if they showed specific enrichment of SPDs within conserved regions. When clades A, B, C, D and F were analysed together, the pattern of clustering of SPDs was highly significant ($P < 0.005$) allowing us to reject the null hypothesis that they were randomly distributed across the genome. SPDs were also significantly clustered in conserved regions when separately analysing clade C ($P \leq 0.005$), clade D ($P \leq 0.05$), and clade F ($P \leq 0.000001$), while clade B was suggestive ($P \leq 0.08$) and clade A was not significant ($P \leq 0.2$). The failure to observe a significant clustering of SPDs in clade A may be due to its considerable substructure that suggests it may actually be comprised of two or more distinct groups. Nonetheless, it is clear that SPDs are often clustered and are found with increased frequency in conserved, shared regions of the genome. SPDs within these regions also share the recent ancestry of the surrounding conserved regions when analysed using phylogenetic trees (Supplementary Figure 17). These findings are consistent with the hypothesis that recent inheritance of conserved blocks location (Supplementary Fig. 12). To compare the ancestry of different regions of the genomes, we partitioned the genome into two segments based on regions that exhibited low SNP diversity in at least one clade (defined as the union of all regions that were ‘conserved’ in at least one clade) versus regions that showed high SNP diversity in all clades (defined as ‘non-conserved’) (Supplementary Data 9 and 10). We then compared unrooted phylogenetic trees for the conserved versus non-conserved regions using a Robinson–Foulds distance metric, which measures the degree of difference, or distance, between the two sets of trees. This analysis revealed that the ancestry of the conserved haploblocks was significantly different than that of the non-conserved regions for all 14 chromosomes (Fig. 5b). In addition, we generated neighbour networks based on the conserved versus non-conserved regions (Supplementary Fig. 16). The conserved region network most closely resembled the total SNP network (Fig. 4a) and the network based on the non-conserved regions grouped most strains in similar clades, with several notable exceptions (Supplementary Fig. 16). These findings illustrate the importance of the conserved blocks in influencing the grouping of genotypes into clades.

Figure 5 | Analysis of conserved regions within and between clades. (a) Average number of pairwise SNPs that are shared by members of clade A and clade D, y axis 0–60 (outliers >60 not plotted). Each data point represents the average per 10 kb window. Red lines indicate chromosome boundaries. (b) Standardized Robinson–Foulds (RF) distances between phylogenetic trees for conserved and non-conserved regions of the genome, based on average pairwise SNP diversity within each clade compared by maximum likelihood. The standardized RF distance equals the proportion of negative branches between the two trees. Threshold distances for statistical significance ($P \leq 0.002$ per chromosome or $\leq 0.028$ for genome wide) were based on 500 bootstrap trees that were used to establish the variation in tree estimates (the height of the threshold bar represents the 99.998% confidence interval of the tree distance). (c) Distribution of conserved regions shared by clades. Horizontal bars below each clade represent the SNP diversity along the combined chromosomes (Chrs). The bar is divided into conserved regions, where the average pairwise SNP diversity is low (red, $\leq 3$ SNPs per 10 kb window), versus non-conserved regions where the SNP diversity is high (grey, $>3$ SNPs per 10 kb window) based on average pairwise comparisons. Above the bar, the distribution of SPDs within conserved regions for each clade are shown as different colour circles for each gene family: TgFAM (purple), SRS (white), ROP (blue), MIC (green) and GRA (yellow). Circle sizes are proportional to the number of genes, reflecting tandem clusters or CNV. SPDs that are shared across more than one clade contain a vertical black line and the names are shown in the top-most clade in which they occur. Grey lines indicate chromosome boundaries.
containing specific SPDs is associated with successful expansion of specific lineages and suggest that SPDs impart a selective advantage to members of specific clades.

The co-inheritance of SPDs within conserved regions provides a tentative list of candidates for further study of genes that may underlie important biological traits shared by specific clades. Among these are a number of SPDs previously implicated in acute virulence in the mouse: for example, ROP17 (ref. 44) found in conserved regions in clades A and B, ROP5 (ref. 29) found in conserved regions in clades A and C, and GRA3 (ref. 45) found in conserved regions in A, C, D, and F (Fig. 5c). Low diversity regions also contain a number of SRS genes, encoding immunologically dominant surface proteins, which have previously been implicated in host cell invasion including sporozoite SAG (Sp-SAG also known as SRS28)46 found in conserved regions in clades A and C, and SAG2A (SRS34A) (M.E.G., unpublished) found in conserved regions in clades C and D (Fig. 5c, Supplementary Data 1). A second pattern that emerges from this analysis is the presence of clusters of SPDs that are clade-specific, for example clusters of SRS and TgFAM genes on chromosomes IV, V and IX found in clade C, and clusters of various SPDs on chromosomes II, IX, XI and XII in clade D (Fig. 5c, Supplementary Data 11). Although the specific roles of these genes are unknown, they may underlie common traits that distinguish phenotypes characteristic of specific clades.

Discussion

Toxoplasma gondii belongs to a diverse and ancient phylum of parasites that antedates the wide range of vertebrate hosts that they currently inhabit. It shares a core set of genes and metabolic processes with closely related tissue-cyst forming coccidian parasites. Despite having similar genomic content, these organisms differ dramatically in their host range, pathogenicity and modes of transmission. We demonstrate here that T. gondii is demarcated from its closest relatives by the expansion of parasite-specific SPDs that are involved in host–pathogen interactions. Diversification of SPDs also highlights key differences among major clades of T. gondii, which are distinguished by common inheritance of large haploblocks in their genomes. Shared inheritance of large haploblocks among related strains reinforces the hypothesis that recombination in the wild, while infrequent, drives important biological adaptations48,49. The distribution of clustered SPDs within conserved regions that show common ancestry identifies a number of candidate genes that may influence both clade specific and more broadly shared traits. Overall, the phenotypic traits of individual strains are likely determined by both their core ancestral genomes, and inheritance of conserved haploblocks, which together comprise their mosaic genomes.

The mosaic genomic patterns seen in specific clades may underlie differences in population structure that exist in different T. gondii populations between North and South America48. Although the common ancestry of conserved blocks among otherwise different genotypes is consistent with recent introgression, the conservation of these regions among members of a given clade may reflect several different mechanisms. Such shared haploblocks may be retained in the face of ongoing recombination in outbreeding populations, suggesting they impart a selective advantage. Alternatively, they may simply reflect recent admixture that has not been eroded due to infrequent recombination, such as in clonal populations. Regardless of their exact histories, strains that inherit conserved haplotype blocks in common will also share clusters of highly related genes, including SPDs that may influence traits such as transmission, host range and pathogenesis.

Expansion of polymorphic genes that are important in pathogenicity is also a key feature of other pathogen genomes. One feature they share in common is that the amplified genes typically encode surface or secretory proteins that interact directly with the host, either to mediate attachment or immune evasion. Examples include: the expansion of surface antigen variants encoded by VAR genes in the Plasmodium falciparum genome50 and the unrelated yet expanded VIR genes in Plasmodium vivax51, variant surface glycoprotein encoding genes (VSG) in Trypanosoma brucei52, and the expansion of RXLR effectors in oomycetes53. It is noteworthy that while gene expansion and diversification are common to each of these examples, the protein families involved are largely distinct and reflect the specialized biology of these diverse pathogens. This pattern suggests that expansion of polymorphic gene families is a common theme that underlies important changes in host range and transmission that characterize the evolution of pathogens in their diverse hosts.

Methods

Propagation of strains and isolation of gDNA. Sixty-two representative strains of T. gondii were selected from different haplogroups from around the world (Supplementary Data 1)46. Strains were cultured in human foreskin fibroblast cells, as described previously16.

Genome sequencing of T. gondii reference strains. Sequencing of 16 reference strains of T. gondii, and of one isolate of H. hammondi, was conducted using a combination of 454, and Illumina PE sequencing technologies (Supplementary Table 1). Sequence reads were screened for contamination and reassembled using Celera Assembler software44 or Newbler v2.6 (ref. 54). Scaffolds were then aligned with MUMmer55 to T. gondii ME49 chromosome sequences from ToxoDB v8.0 (http://ToxoDB.org) to generate super-scaffolds spanning entire chromosomes.

Annotated genomes were deposited into National Center for Biotechnology (NCBI).

Genome sequencing for SNP discovery. For each of the remaining 46 non-reference strain (Supplementary Data 1), a single Illumina PE barcoded library was prepared from tachyzoite gDNA. Libraries were then pooled into groups of nine samples and sequenced multiplexed in a single lane of an Illumina HiSeq 2000 machine. Sequencing reads were deposited in the Sequence Read Archive repository at NCBI.

Sequencing of tachyzoite messenger RNA samples. To aid in the curation of ME49 gene models, two tachyzoite-specific Illumina complementary DNA libraries were constructed from mRNA isolated from tachyzoite cultures from ME49 and GT1 strains. Each library was then sequenced in a single lane of an Illumina Genome Analyzer II machine.

Structural and functional annotation of the ME49 genome. Gene annotations were derived by comparison of the existing ME49 reference genome (http://ToxoDB.org) using a combination of evidence from RNAseq data, cDNA/EST sequences and a variety of software tools to predict potential protein-coding genes using an in-house pipeline at J. Craig Venter Institute (JCVI) (Supplementary Methods). Predicted proteins were run through JCVI’s auto-naming pipeline, that assign product names based on a number of sequence similarity searches including blastp searches against the previous T. gondii ME49 proteome (ToxoDB v8.0; http://ToxoDB.org) and the GenBank non-redundant protein database, HMM searches against Pfam and TIGRfam56 databases, and RPS-Blast searches against the NCBI-CDD database56. Proteins without any significant hit to other proteins or protein domains were flagged as ‘hypothetical protein’. The final list of product names was then curated by researchers from the Toxoplasma research community before being assigned to working models. ME49 protein-coding genes were assigned similar pub_locus identifiers to the previous genome assembly while newly predicted protein-coding genes were assigned completely new pub_locus identifiers (Supplementary Methods).

Annotation of T. gondii reference strains and H. hammondi. Functional annotation of protein-coding genes in other T. gondii reference strains was performed as above. Genes syntenic to ME49 inherited their product names, GO terms, and Enzyme Commission numbers, while non-syntenic genes acquired their names and other functional annotations from the output of JCVI’s autonaming pipeline. Structural and functional annotations of H. hammondi were carried out following a similar approach with slight modifications (Supplementary Methods).
Domain Identification of T. gondii novel gene families. To identify known protein domains the T. gondii ME49 proteome was searched against Pfam and TIGRFam HMM profiles using HMMER3 (ref. 57). Peptides matching a particular HMM profile were assigned to that domain and remaining peptides searched against each other using blastp to identify potential novel domains. The top five protein families containing novel para domains (TgFAMs A to E) were analysed using Phobius to identify signal peptides and transmembrane domains. De novo identification of conserved protein domains across members of the same gene family was carried out with MEME Expression levels for the TgFAMs were calculated using a custom script over a 10-kb moving window and positions with informative base calls for all 62 strains were identified, generating a final list of 802,764 SNPs that were used for analysis.

SNP analysis. Illumina reads for each of the 61 other genomes were aligned using Bowtie2—end-to-end50 against the ME49 reference genome assembly (release date 23 April 2013), identifying a total of 2,342,433 SNPs across all strains. Positions with informative base calls for all 62 strains were identified, generating a final list of 802,764 SNPs that were used for analysis.
18. Blazewiński, T. et al. Systems based analysis of the Sarcozystis neurona genome identifies pathways that contribute to a heteroxenous life cycle. MBio 6, 02445–14 (2015).

19. Song, C. et al. Comparative genomics of the apicomplexan parasites Toxoplasma gondii and Neospora caniunum: coccidia differing in host range and transmission strategy. PLoS Pathog. 8, e1002567 (2012).

20. Su, C. L. et al. Globally diverse Toxoplasma gondii isolates comprise six major clades originating from a small number of distinct ancestral lineages. Proc. Natl Acad. Sci. USA 109, 5844–5849 (2012).

21. Khan, A. et al. Composite genome map and recombination parameters derived from three archetypal lineages of Toxoplasma gondii. Nucleic Acids Res. 33, 2980–2992 (2005).

22. Billker, O., Lourido, S. & Sibley, L. D. Calcium-dependent signaling and kinases in apicomplexan parasites. Cell Host Microbe 5, 612–622 (2009).

23. Balaji, S., Babu, M. M., Iyer, L. M. & Aravind, L. Discovery of the principal specific transcription factors of Apicomplexa and their implication for the evolution of the AP2-integrate DNA binding domains. Nucleic Acids Res. 33, 3995–4006 (2005).

24. Wasmuth, J. D. et al. Integrated bioinformatic and targeted depletion analyses of the SRS gene superfamily identify SRS29C as a negative regulator of Toxoplasma virulence. mBio 3, e00321–12 (2012).

25. Arredondo, S. A. et al. Structure of the Plasmodium 6-cysteine 448/45 domain. Proc. Natl Acad. Sci. USA 109, 6692–6697 (2012).

26. Tonkin, L. M. et al. Structural and biochemical characterization of Plasmodium falciparum 12 (Pf12) reveals a unique interdomain organization and the potential for an antiparallel arrangement with Pf41. J. Biol. Chem. 288, 12805–12817 (2013).

27. Song, C. et al. Metabolic reconstruction identifies strain-specific regulation of virulence in Toxoplasma gondii. Mol. Syst. Biol. 9, 708 (2013).

28. Debary, J. D. & Kissing, J. C. Jumbled genomes: missing Apicomplexan genomes from three archetypal lineages of Plasmodium falciparum. PLoS Pathog. 10, (2014).

29. Behnke, M. S. et al. Virulence differences in Toxoplasma mediated by amplification of a family of polymorphic pseudoexonases. Proc. Natl Acad. Sci. USA 108, 9631–9636 (2011).

30. Taylor, S. et al. A secreted serine-threonine kinase determines virulence in the eukarystic pathogen Toxoplasma gondii. Science 314, 1776–1780 (2006).

31. Selmecki, A. M. et al. Polyplody can drive rapid adaptation in yeast. Nature 519, 349–352 (2015).

32. Adomako-Ankomah, Y., Wier, G. M., Borges, A. L., Wand, H. E. & Boyle, J. P. Differential locus expansion distinguishes Plasmodiophorinae species and closely related strains of Toxoplasma gondii. mBio 5, e01003–e01013 (2014).

33. Cotterill, C. R., Matthews, S. & Tomley, F. The molecular basis for the distinct host and tissue tropisms of coccidian parasites. Mol. Biochem. Parasitol. 186, 1–10 (2012).

34. Dalmasso, M. C., Carmona, S. J., Angel, S. O. & Aguero, F. Characterization of Toxoplasma gondii subtropical-region like: identification of a long-range compositional bias that is also associated with gene-poor regions. BMC Genomics 15, 350 (2014).

35. Scherf, A., Lopez-Rubio, E. J. I. & Riviere, L. Antigenic variation in Trypanosoma cruzi: a comparative analysis with Trypanosoma brucei. Bioinformatics. 25, 2078–2079 (2009).

36. Behnke, M. S., Zhang, T. P., Dubey, J. P. & Sibley, L. D. Toxoplasma gondii merozoite gene expression analysis with comparison to the life cycle discloses a unique expression state during enteric development. BMC Genomics 15, 350 (2014).

37. Khan, A. et al. Recent transcriptional sweep of Toxoplasma gondii driven by a single monomorphic chromosome. Proc. Natl Acad. Sci. USA 104, 14872–14877 (2007).

38. Minot, S. et al. Admixture and recombination among Toxoplasma gondii lineages explain global genome diversity. Proc. Natl Acad. Sci. USA 109, 13458–13463 (2012).

39. Huson, D. H. & Bryant, D. Application of phylogenetic networks in evolutionary studies. Mol. Biol. Evol. 23, 254–267 (2006).

40. Alexander, D. H., Novembre, J. & Lange, K. Fast model-based estimation of adaptive ancestry in unrelated individuals. Genome Res. 19, 1655–1664 (2009).

41. Zhao, P. et al. Genetic characterization of Toxoplasma gondii isolates from pigs in China. J. Parasitol. 96, 1027–1029 (2010).

42. Khan, A. et al. Geographic separation of domestic and wild strains of Toxoplasma gondii in French Guiana correlates with a monomorphic version of chromosome1A. Plos Negl. Trop. Dis. 8, e3182 (2014).
Author contributions
H.L., M.H. and S.K. performed genome assembly and/or annotation; H.L., A.K., M.S.B., S.N., L.S.S., D.P., R.P.B., M.A.D.-M., B.D.G., L.L. and J.C.K. generated data files for analyses; H.L., A.K., M.S.B., S.N., L.S.S., D.K.H., M.E.G., I.P., L.L., J.C.K., D.S.R. and L.D.S. analysed data and generated Figures or Tables; D.A., M.L.D., J.P.D., H.M.F., S.M.G., M.E.G., C.S. and X.Q.Z. provided key biological materials or data; J.W.A., J.P.B., J.C.B., M.E.G., K.K., B.M.R., M.W.W., J.P., J.C.K., D.S.R. and L.D.S.; provided advice on project design and analyses; H.L., J.P., J.C.K., D.S.R. and L.D.S. supervised the project; H.L., M.S.B., B.M.R., J.P., L.L., J.C.K. and L.D.S. wrote the paper, with input from all authors.

Additional information
Accession codes: Assembled and annotated sequences for the genomes of the *T. gondii* reference strains have been deposited in GenBank at NCBI under accession codes GT1 (GCA_000149715.2), VEG (GCA_000150015.2), ME49 (GCA_000006565.2), MAS (GCA_000224865.2), RUB (GCA_000224805.2), CAST (GCA_000256705.1), TgCBr5 (GCA_000259835.1), P89 (GCA_000224885.2), VAND (GCA_000224845.2), COUG (GCA_000338675.1), ARI (GCA_000250965.1), TgCrPCr2 (GCA_000256725.1), GAS-2007-GAL-DOM2 (GCA_000325525.2), TgCtCo5 (GCA_000278365.1), FOU (GCA_000224905.2) and TgCtBr9 (GCA_000224825.1). Assembled and annotated sequences for the genomes of the *H. hammondi* reference strains have been deposited in GenBank at NCBI under accession code GCA_000258005.2. The annotation of reference genomes are also available at ToxoDB (http://toxodb.org/common/downloads/White-Paper/ProvisionalGenomes/). RNA-Seq reads for *T. gondii* strains have been deposited in the Sequence Read Archive at NCBI under accession codes GT1 (SRX099798) and ME49 (SRX099799). Sequence reads from the remaining *T. gondii* strains have been deposited in the Sequence Read Archive at NCBI under the following accession codes RH-88 (SRX160127, SRX160126), RH-JSR (SRX159850, SRX159851), TgCrCr1 (SRX160807), SOU (SRX160119), TgCrKz2 (SRX099785), CASTELLS (SRX099789), TgCatBr1 (SRX099790), TgCatBr18 (SRX099794), TgCatBr34 (SRX160051), BRC TgH 18002 GUY-KOE (SRX099786), BRC TgH 18003 GUY-MAT (SRX099783), GUY-2004-ABE (SRX160132), TgBrCr1 (SRX160143), TgCrBr141 (SRX160124), TgCatBr25 (SRX160134), TgCatBr10 (SRX099791), TgCatBr64 (SRX160743), BRC TgH 18001 GUY-DOS (SRX099782), TgCrCr10 (SRX099784), B41 (SRX099774), RAY (SRX099793), TgCatBr44 (SRX160141), GAB1-2007-GAL-DOM10 (SRX159841, SRX159839), GAB3-2007-GAL-DOM9 (SRX160125), GAB5-2007-GAL-DOM6 (SRX159842, SRX159840), TgCatPrCr3 (SRX09979), BRC TgH 18009 (SRX171132), TgH18 18021 (SRX160041), BRC TgH 20005 (SRX157499, SRX157465), BRC TgH 21016 (SRX099775), GUY-2004-JAG1 (SRX099776), B73 (SRX159844, SRX159843), PRU (SRX099792), M7741 (SRX159890, SRX159849), ROD (SRX160128, SRX160127), TgShUS28 (SRX160130), BOF (SRX099774), GAB5-2007-GAL-DOM1 (SRX160069), TgCatBr26 (SRX099780), TgCatBr72 (SRX160049), TgDqCt17 (SRX099787), TgCatBr15 (SRX099779), G662M (SRX160052), GAB3-2007-GAL-DOM2 (SRX160123), TgCatBr3 (SRX160142) and TgH26044 (SRX160050).

Supplementary Information accompanies this paper at http://www.nature.com/naturecommunications

Competing financial interests: The authors declare no competing financial interests.

Reprints and permission information is available online at http://npg.nature.com/reprintsandpermissions/

How to cite this article: Lorenzi, H. et al. Local admixture of amplified and diversified secreted pathogenesis determinants shapes mosaic *Toxoplasma gondii* genomes. *Nat. Commun.* 7:10147 doi: 10.1038/ncomms10147 (2016).

This work is licensed under a Creative Commons Attribution 4.0 International License. The images or other third party material in this article are included in the article’s Creative Commons license, unless indicated otherwise in the credit line; if the material is not included under the Creative Commons license, users will need to obtain permission from the license holder to reproduce the material. To view a copy of this license, visit http://creativecommons.org/licenses/by/4.0/