nature COMMUNICATIONS

ARTICLE

Received 21 Jan 2015 | Accepted 9 Nov 2015 | Published 7 Jan 2016

DOI: 10.1038/ncomms10147

OPEN

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

Hernan Lorenzi[1,*], Asis Khan[2,3,*], Michael S. Behnke[2,4,*], Sivaranjani Namasivayam[5,6], Lakshmipuram S. Swapna[7,8], Michalis Hadjithomas[1,†], Svetlana Karamycheva[1], Deborah Pinney[9], Brian P. Brunk[9], James W. Ajioka[10], Daniel Ajzenberg[11], John C. Boothroyd[12], Jon P. Boyle[13], Marie L. Dardé[11], Maria A. Diaz-Miranda[9], Jitender P. Dubey[14], Heather M. Fritz[15], Solange M. Gennari[16], Brian D. Gregory[9], Kami Kim[17], Jeroen P.J. Saeij[18], Chunlei Su[19], Michael W. White[20], Xing-Quan Zhu[21], Daniel K. Howe[22], Benjamin M. Rosenthal[14], Michael E. Grigg[3], John Parkinson[7,8], Liang Liu[23,24], Jessica C. Kissinger[5,6,24], David S. Roos[9] & L. David Sibley[2]

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. We 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 Rougea Louisian 70803, USA. [5] Department of Genetics, University of Georgia, Athens, Georgia 30602, USA. [6] Center for Tropical and Emerging Global Diseases, University of Georgia, Athens Georgia 30602, USA. [7] Program in 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 Biology, 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] Departments 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, David, 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 (CDDI), 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 ~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 enjoys worldwide prevalence in animals including humans[5]. Infections with *T. gondii* are common[6], yet they typically only cause disease in immunocompromised hosts, or as a result of transplacental infection[7]. *T. gondii* is equipped with excellent forward and reverse genetic tools, providing a model for many less-tractable apicomplexan parasites[8]. 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[9], 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.

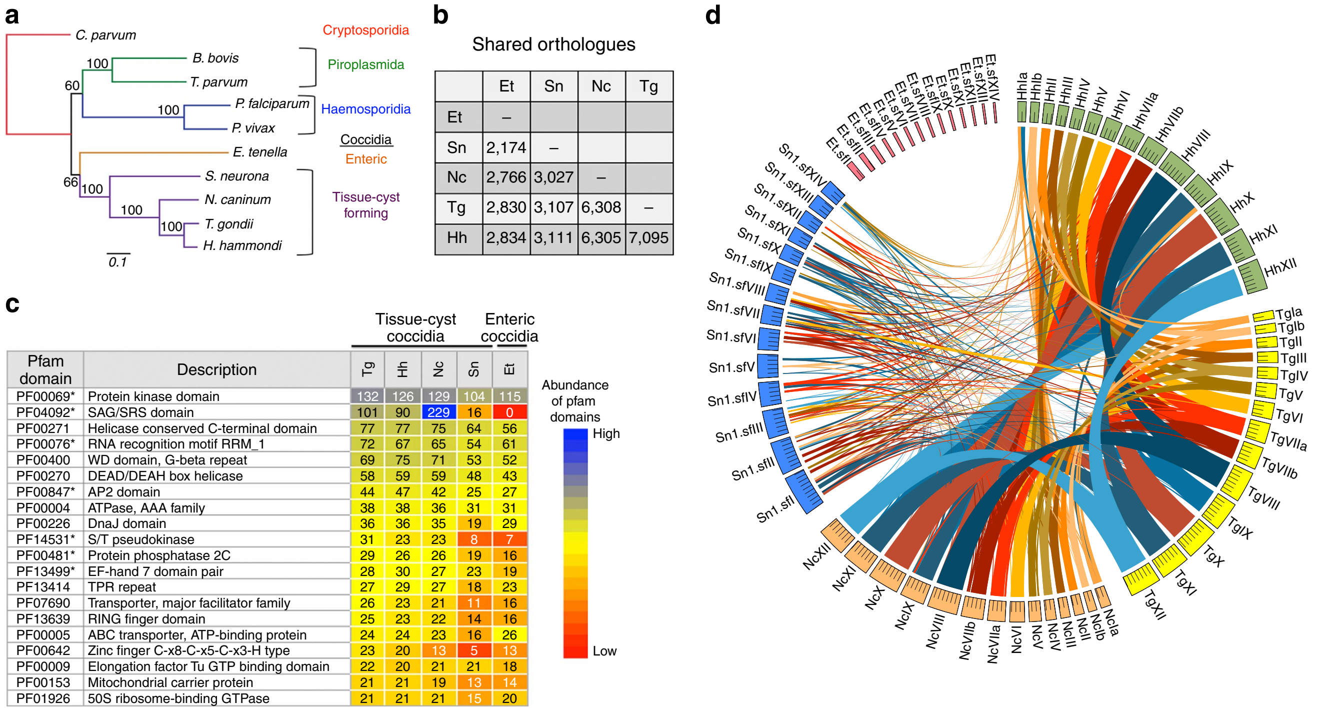
In North America and Europe, the population structure of *T. gondii* is dominated by three prevalent clonal lineages[10], which coexist with much more rare, genetically diverse isolates. A fourth clonal lineage is largely confined to North America, where it is more common in wild animals[11]. In contrast, 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 × 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 × 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 × average sequence coverage) (Supplementary Table 1). The remaining strains were sequenced using Illumina only (~42 × 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



Shared orthologues

	Et	Sn	Nc	Tg
Et	–			
Sn	2,174	–		
Nc	2,766	3,027	–	
Tg	2,830	3,107	6,308	–
Hh	2,834	3,111	6,305	7,095

c

Pfam domain	Description	Tissue-cyst coccidia Tg	Hh	Nc	Sn	Enteric coccidia Et
PF00069*	Protein kinase domain	132	126	129	104	115
PF04092*	SAG/SRS domain	101	90	229	16	0
PF00271	Helicase conserved C-terminal domain	77	77	75	64	56
PF00076*	RNA recognition motif RRM_1	72	67	65	54	61
PF00400	WD domain, G-beta repeat	69	75	71	53	52
PF00270	DEAD/DEAH box helicase	58	59	59	48	43
PF00847*	AP2 domain	44	47	42	25	27
PF00004	ATPase, AAA family	38	38	36	31	31
PF00226	DnaJ domain	36	36	35	19	29
PF14531*	S/T pseudokinase	31	23	23	8	7
PF00481*	Protein phosphatase 2C	29	26	26	19	16
PF13499*	EF-hand 7 domain pair	28	30	27	23	19
PF13414	TPR repeat	27	29	27	18	23
PF07690	Transporter, major facilitator family	26	23	21	11	16
PF13639	RING finger domain	25	23	22	14	16
PF00005	ABC transporter, ATP-binding protein	24	24	23	16	26
PF00642	Zinc finger C-x8-C-x5-C-x3-H type	23	20	13	5	13
PF00009	Elongation factor Tu GTP binding domain	22	20	21	21	18
PF00153	Mitochondrial carrier protein	21	21	19	13	14
PF01926	50S ribosome-binding GTPase	21	21	21	15	20

Abundance of pfam domains
High
Low

Figure 1 | Comparative genomics of tissue-cyst forming coccidian parasites. (**a**) Phylogenetic tree of selected apicomplexans based on a conserved DEAD box helicase protein (TGME49_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	Myeloencephalitis in horses and marine mammals	None
Neospora caninum	Canine	Bovine	Vertical asexual or sexual–asexual cycle	Congenital, abortion in cattle neurological, paralysis in dogs	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

(GO) 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 kinases[21], such as the expanded polymorphic family of rhoptry (ROP) kinase virulence determinants in *T. gondii*[12] and of calcium-binding motifs, including within a family of calcium-dependent protein kinases[22]. However, the abundance of RNA-binding proteins (RMR and

Table 2 | Summary of genome features for *T. gondii* and representative apicomplexans.

Feature	*S. neurona*[*]	*N. caninum*[†]	*H. hammondi*[‡]	*T. gondii* ME49[§]
Estimated size	~127 Mb[*]	~62 Mb[†]	~65 Mb	~65 Mb
Assembly length without sequencing gaps (bp)	117,871,271	57,524,119	67,460,985	65,464,221
Number of scaffolds[‖]	116	NA[¶]	99	47
Scaffolds N50 (bp)	2,890,735	NA	1,494,935	6,301,488
Number of contigs[#]	8,903	241	1,337	410
Contigs N50 (bp)	20,915	405,161	84,429	1,219,553
Sequencing depth	375 ×	8 × [†]	66 ×	26.5 ×
# Chromosomes	NA	14[†]	14	14
# Protein-coding genes	7,093	6,936[**]	8,004[‡]	8,322[§]
GC content	51.5%	54.8%	52.5%	52.2%
% Protein-coding sequence[††]	50.9%	59%	57.3%	60.5%
Average length of protein-coding genes[‡‡](bp)	9,121	4,892	4,868	4,778
Average number of exons per protein-coding gene	5.5	12	11.7	11.5

NA, not available.
*All sequence reads were deposited in the National Center for Biotechnology (NCBI) sequence read archive under accession SUB554996 ref. 14.
†Reid *et al.*[15].
‡GenBank Assembly ID GCA_000258005.2.
§GenBank Assembly ID GCA_000006565.2.
‖Scaffolds >10,000 bp.
¶960 Scaffolds, any size.
#Contigs >2,000 bp.
**GenBank Assembly ID GCA_000208865.2.
††Exons and introns, without UTRs.
‡‡Without UTRs.

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 in tissue-cyst forming coccidian parasites (Fig. 1c), consistent with these being major transcription factors in apicomplexans[23]. Also prevalent in *N. caninum*, *T. gondii* and *H. hammondi* are a family of surface antigens (SAG) called the SRS family (Fig. 1c), which are amplified and highly divergent among tissue-cyst forming coccidian parasites[24]. These appear less abundant in *S. neurona* and absent in *E. tenella* (Fig. 1c), 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[18] and a related family is found in *Plasmodium*[25]. SRS proteins and related 6-Cys proteins share a common extracellular structural domain[26], are typically GPI-anchored, and are thought to play diverse roles in cell attachment, invasion and development.

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[15]. 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 (Supplementary Fig. 2). 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[27]. 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 (Supplementary Fig. 2, Supplementary Data 3).

Finally, we compared the position of genes across the chromosomes to establish the extent of synteny (Fig. 1d, Supplementary Fig. 3, Supplementary Table 2). 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*[15,28] (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 (Fig. 1d). 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*[18] (Fig. 1d). 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[28].

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[29,30]. 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[31].

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[15,32]. 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,

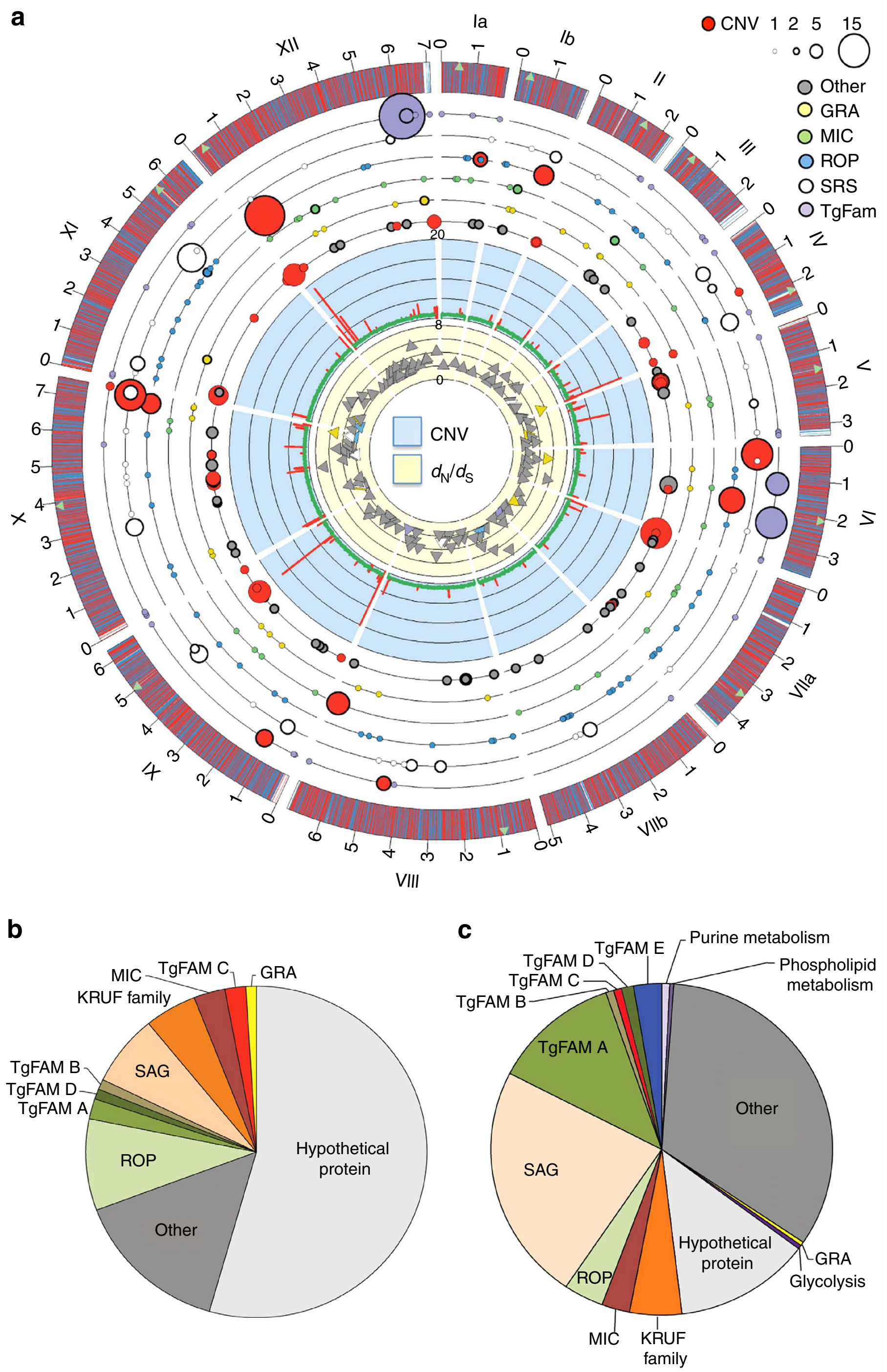


Figure 2 | Composition of the *T. gondii* genome. (**a**) Circos representational plot of the *T. gondii* genome based on the reference strain ME49. The outside track plots the position of genes on the 14 chromosomes (numbered outside in Roman numerals with size intervals given in Arabic numerals): top strand gene (blue), bottom strand gene (red), centromere locations (green triangles). The innermost track with yellow background is a scatter plot of genes with d_N/d_S ratio ≥ 2 (136 genes): *y* axis 0–8, colours indicate gene family. The second innermost track with blue background is a histogram of regions with copy number variation (CNV) in ME49 using a rolling window of 2,500 bp: *y* axis 0–20, 1X copy level (green), region with CNV (red). The next six circular tracks indicate gene families with tandem duplicates or CNV. From the inside to outside, these six levels represent the following categories: other (grey), *GRA* (yellow), *MIC* (green), *ROP* (blue), *SRS* (white) and *TgFAM* (purple) genes. The size of the circle is relative to the number of genes (range 1–15). Circles with a thick border indicate tandem arrays while red circles indicate CNV. (**b**) Frequency of different gene families among genes with CNV includes SPDs such as *SAG*, *ROP*, *MIC* as well as *TgFAM* genes. (**c**) Frequency of different gene families among genes with tandem duplications in the assembled genomes includes SPDs such as *SAG*, *ROP*, *MIC* and *TgFAM* 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)[33], modification of host immunity (GRA and ROP proteins)[12] or adherence and immune evasion (SRS)[24]. 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[32]. 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 (d_N) versus synonymous (d_S) mutations (Fig. 3a). Among these, the *GRA*, *ROP* and *SAG* genes show some of the highest levels of d_N/d_S, while metabolic enzymes typically show selection for conservation, as seen by low levels of d_N/d_S (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,

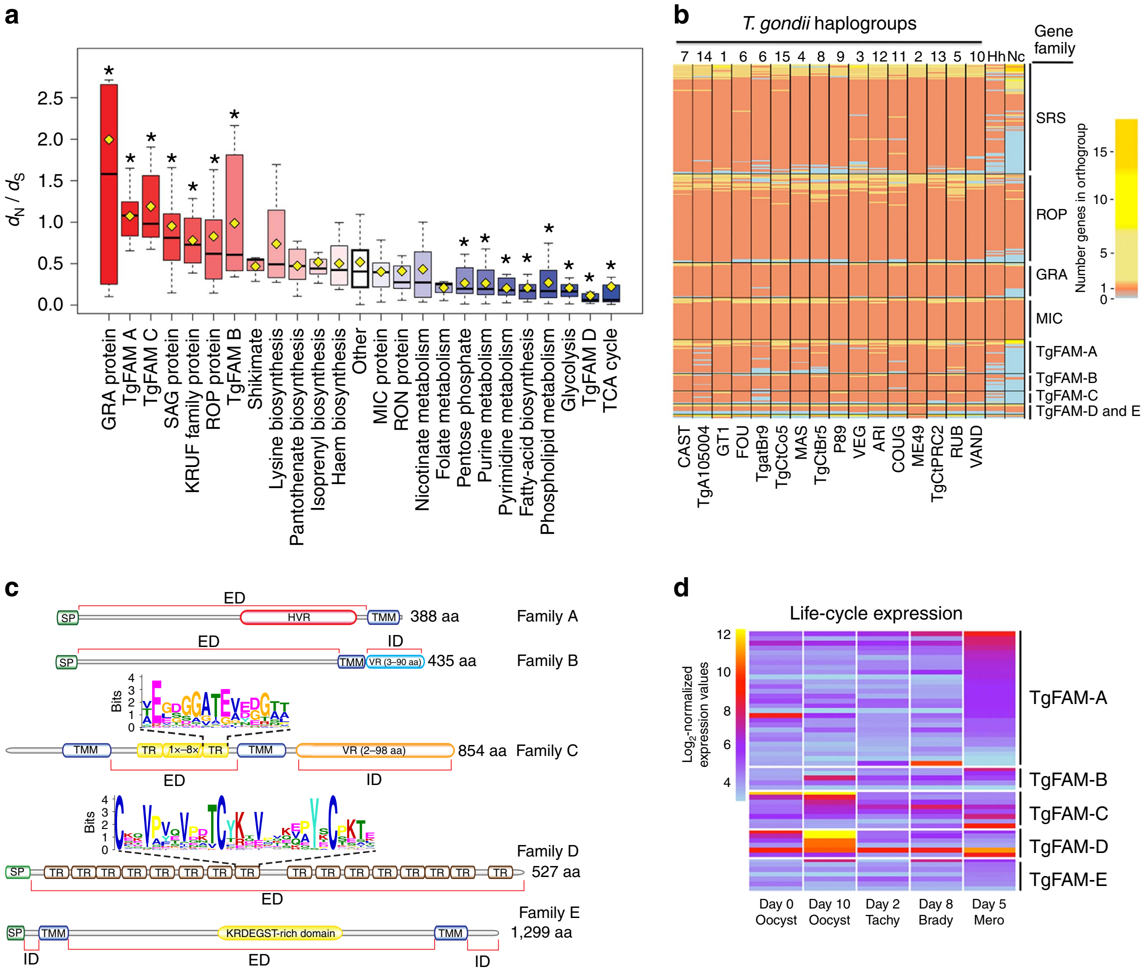


Figure 3 | Analysis of expanded or positively selected gene families in *T. gondii*. (**a**) Analysis of positive selective pressure among *T. gondii* gene families based on d_N/d_S. Red values indicate categories with significantly elevated d_N/d_S ratios, indicating positive selection for diversification, 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 d_N/d_S median; yellow square, d_N/d_S mean; mean values significantly different than the category 'Other' ($P \leq 0.05$, 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. caninum*, 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 $\log_2$-normalized microarray expression values for *TgFAM* genes across the *T. gondii* life cycle[36,66,67]. Background values for the array are near $\log_2$ 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. hammondi* 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. hammondi 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. hammondi*, versus the pairwise comparison between the later two species. Among these distantly 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. hammondi* 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. 8 c–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*[18]), 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* family[34]. 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 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. hammondi* and *N. caninum* (Fig. 3b, Supplementary Data 7). In particular, *H. hammondi* and *N. caninum* contain far fewer members of *TgFAMA* 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)[36], 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 haplogroups[16,37,38], 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 network[39] 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)[16]. Importantly, similar groupings were defined using admixture[40] (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 previously[16,37].

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 ($\sim$60%), interspersed with regions that are divergent (Fig. 4b), consistent with previous findings that these two groups are closely related[11]. In contrast, TgCtPRC2 (haplogroup 13), which is a common clonal genotype in China[41], shares fewer regions with ME49 ($\sim$40%) and COUG (haplogroup 11), which represents a rare North American lineage found in wild animals, showed almost no conserved regions with ME49 ($<$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 Ia, 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 cats[42].

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

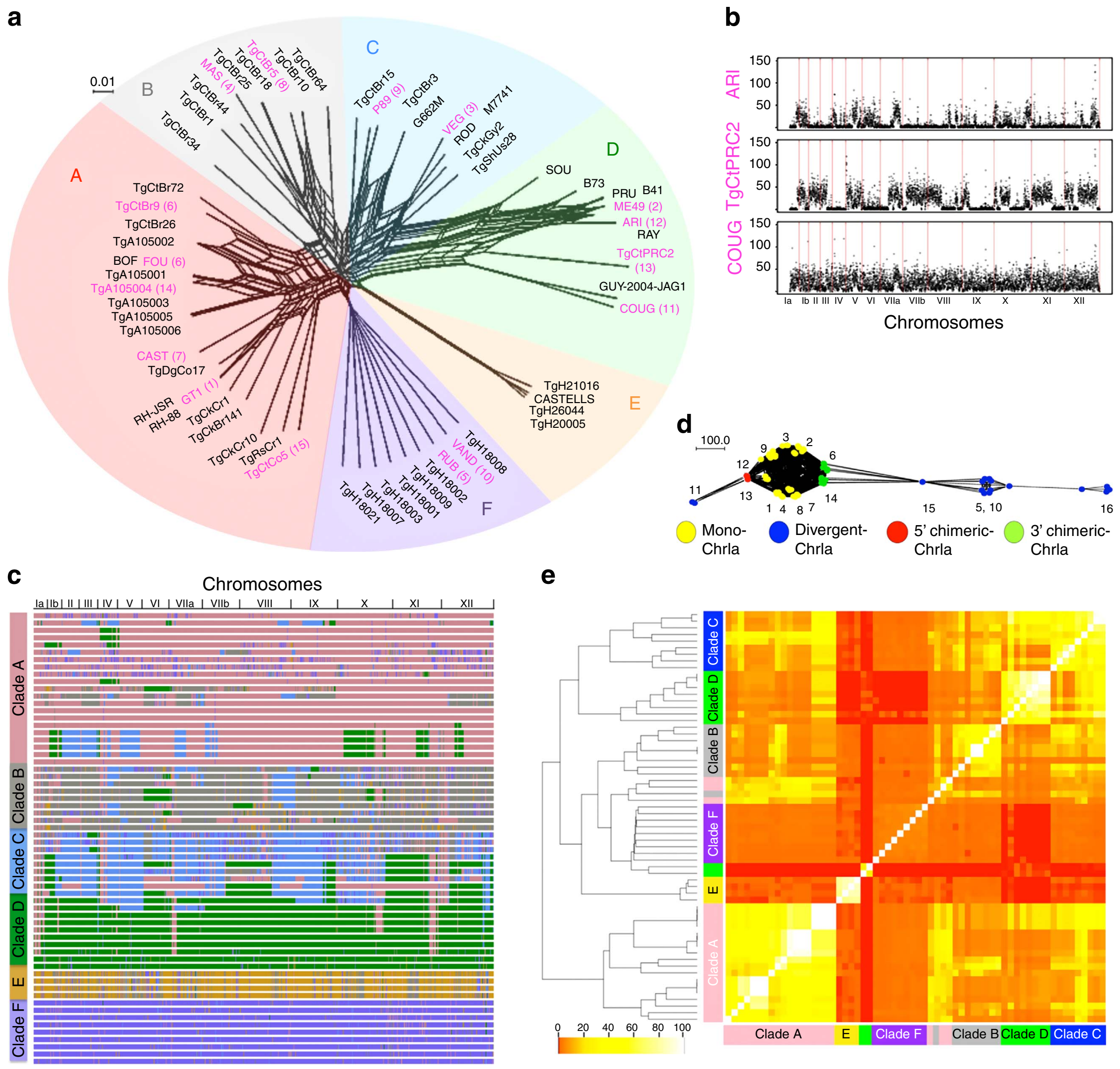


Figure 4 | Comparative genomics and population structure of *T. gondii*. (**a**) Population genetic structure of *T. gondii*. Neighbour-net analysis based on genome-wide SNPs (802,764 common data points) from 62 isolates of *T. gondii*. Colour wheel indicates major clades of *T. gondii*. Haplogroup numbers are indicated within parenthesis based on previous designation. Pink names denote the representative strains. Scale bar, number of SNPs per site. (**b**) Pairwise comparison of SNPs between indicated strains to ME49 shown across 14 chromosomes. *y* axis = number of SNPs/10 kb window. (**c**) Chromosome painting of 62 *T. gondii* strains. Local admixture analyses were conducted on SNP blocks of size 1,000 on each of the 14 chromosomes. For each SNP block, local admixture was used to assign strains to a particular ancestral population. The shared inheritance of blocks across members reveals colour patterns that extend vertically in the plot. For example, several pink regions show strong vertical patterns bifurcating across multiple clades, although the dominant colour is not meant to imply origin. (**d**) Network of chromosome Ia (ChrIa) showing high conservation within most haplogroups (number) and clades. Monomorphic forms (Mono, 3′ Chimeric, 5′ Chimeric) are shared by most lineages, while a few strains are highly divergent. Scale bar, total number of changes. (**e**) Heatmap clustering of co-inheritance of shared blocks. The percentage of shared blocks between two strains was determined for all 62 × 62 pairwise strain comparisons (1,953 non-redundant comparisons): Scale bar, % shared blocks. Hierarchical clustering on percent shared blocks independently grouped the strains by clade.

performed two types of analysis to cluster strains using the SNP data. First, we analysed SNPs using the linkage model of ChromoPainter in FineStructure[43] 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 pairwise comparisons (1,953 unique comparisons), which produced a highly similar clade structure (Fig. 4e, Supplementary Fig. 15). These analyses reveal that the current population structure is defined by recent genomic admixture, 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[38], 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 with those that were more variable. When SNP diversity was averaged for 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

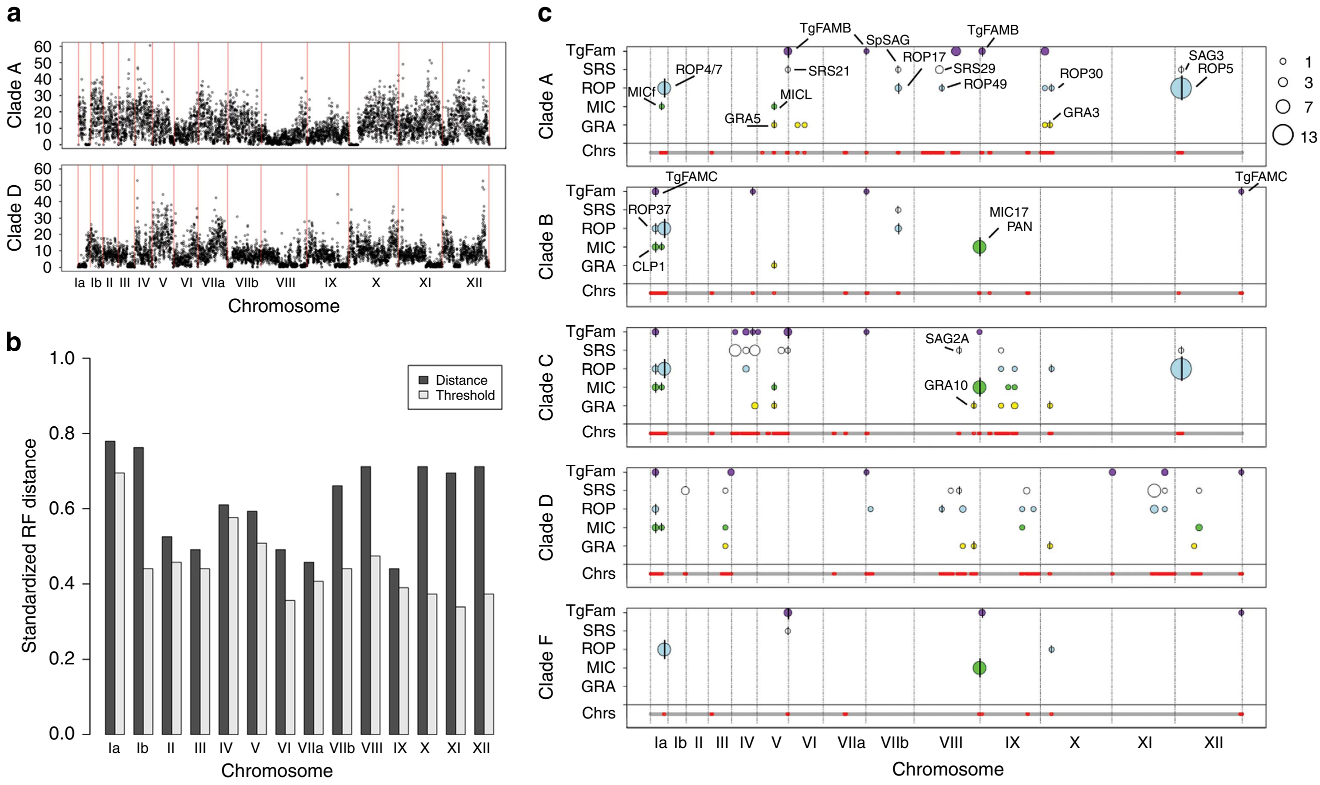


Figure 5 | Analysis of conserved regions within and between clades. (**a**) Average number of pairwise SNPs that are shared by members of clade A and clade D, *y* axis 0–60 (outliers >60 not plotted). Each data point represents the average per 10 kb window. Red lines indicate chromosome boundaries. (**b**) 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 ≤ 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, ≤ 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.

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.

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 ($P \leq 0.005$) 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 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 B, SAG3 (SRS57)[47] 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 11). 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 adaptations[48,49]. 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 America[10]. 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* genome[50], and the unrelated yet expanded *VIR* genes in *Plasmodium vivax*[51], variant surface glycoprotein encoding genes (*VSG*) in *Trypanosoma brucei*[52], and the expansion of RXLR effectors in oomycetes[53]. 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)[16]. Strains were cultured in human foreskin fibroblast cells, as described previously[16].

Genome sequencing of *T. gondii* reference strains. Sequencing of 16 reference strains of *T. gondiii*, 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 software[54] or Newbler v2.6 (ref. 54). Scaffolds were then aligned with MUMmer[55] 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 autonaming 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 TIGRfam[56] databases, and RPS-Blast searches against the NCBI-CDD database[56]. 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). Proteins 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[58] 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[59]. Expression levels for the TgFAMs were obtained from *T. gondii* Affymetrix Array data available from NCBI GEO records GSE32427 and GSE51780.

Estimation of d_N/d_S ratios. Coding sequences from each cluster of orthologous genes from the 16 *T. gondii* reference strains were used to estimate d_N/d_S ratios using a modified version of the Bioperl script *bp_pairwise_kaks.pl* (http://search.cpan.org/dist/BioPerl/scripts/utilities/bp_pairwise_kaks.pl).

SNP identification. Illumina reads for each of the 61 other genomes were aligned using Bowtie2—end-to-end[60] 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.

Analysis of orthologous genes. Annotated proteomes were analysed using OrthoMCL v2.0 (ref. 61) to define orthologous groups. Clusters of orthologous groups were functionally annotated using GO Slim terms, which are designed to group the many different GO terms into smaller groups of related processes (http://geneontology.org/page/go-slim-and-subset-guide). The proteomes were queried against the Pfam HMM database using HMMER3 to estimate the abundance of Pfam domains.

Mapping metabolic differences. SNPs from the 16 *T. gondii* reference strains that correspond to the 382 proteins in the iCS382 metabolic pathway reconstruction of *T. gondii* ME49 (ref. 27) were downloaded from ToxoDB (http://www.toxodb.org/toxo-release4-0/home.jsp).

Network and principal components analyses. Genome-wide SNPs were saved as FASTA files and directly incorporated into SplitsTree v4.4 (ref. 39) to generate unrooted phylogenetic networks using a neighbour-net method and 1,000 bootstrap replicates. Principal components analysis was performed by eigenanalysis of a co-ancestry matrix implemented in fineSTRUCTURE, as described in ref. 43.

Chromosome Ia analysis. SNP data for ChrIa were plotted as a minimum spanning tree using SplitsTree v4.4 (ref. 39) with 2,000 spring-embedded iterations. The 62 strains were clustered into four major groups denoted as monomorphic, divergent, 5′-chimeric and 3′-chimeric chromosome Ia. SNPs present in each cluster were calculated using a custom script over a 10-kb moving window and plotted using Excel.

Admixture analysis. The population genetic structure of *T. gondii* was determined by an unsupervised clustering algorithm, ADMIXTURE[40] with ancestral clusters set from $k = 1$ through 10. The number of ancestral clusters k was determined by estimating the low cross-validation error (CV error) for different k values using five-fold CV.

Co-ancestry heatmap. We developed a co-ancestry heatmap by using the linkage model of ChromoPainter (http://www.paintmychromosomes.com) and fineSTRUCTURE[43] based on the genome-wide SNP data. The burn-in and Markov Chain Monte Carlo (MCMC) after the burn-in were run for 10,000 iterations with default settings.

Estimating CNV. For each strain of *T. gondii*, the respective.sra files were used to align reads to the 14 ME49 reference chromosomes using Bowtie2 with the end-to-end option. The read depth per base pair, or read bases (RB), across 8,320 chromosomal-mapped genes was determined using samtools mpileup[62]. Plots were generated in R (http://www.r-project.org/). *T. gondii* gene families organized in tandem arrays were identified with an in-house *perl* script.

Analysis of OrthoMCL species-specific genes. Genes found to be specific to *T. gondii, H. hammondi or N. caninum* based on OrthoMCL clustering were further analysed using a combination of sequence alignment tools (Supplementary Methods). Genes were classified based on whether they showed a significant ‘Blastp hit’, showed blastn similarity that was either ‘full length’ or constituted an ‘alternative gene model’, or showed no similarity and were ‘unique’. Further analysis was done to investigate these differences by RNA-Seq, analysis of minimum alternative allele frequency and minimum read depth (Supplementary Methods).

Regions of co-inheritance. To determine the extent of recombination and co-inheritance of blocks between strains, low SNP regions (regions of recent co-inheritance or shared blocks) were identified for 10 kb windows for pairwise strain comparisons. A heatmap was generated using the R function heatmap.2 (gplots library (http://www.r-project.org/)) with hierarchical clustering on the % shared blocks value. The number of SNPs per 10 kb window were averaged for all strains within a Clade, and chromosomal regions with low SNP density were identified as above using 10 kb windows that had three or fewer SNPs across a continuous stretch of 10 windows (100 kb), allowing for intermittent outliers.

Identification of SPD genes and clustering within the genome. We identified genes that belong to the SPD families (that is, *MIC*, *GRA*, *ROP*, *SRS* and *TgFAM*) based on the annotation of ME49 accounting for CNV in determining the gene number. We then mapped the position of the SPDs onto the assembled ME49 genome and defined those that fell into conserved or non-conserved regions. To determine if gene type was independent of region type we compared the observed frequency of SPDs and non-SPD genes in conserved versus non-conserved regions of the genome using a χ^2-squared analysis. The null hypothesis was that the distribution would be random, and there would be no difference between observed and expected. A P value of ≤ 0.05 was considered significant cause for rejection of the null hypothesis.

Ancestry of conserved and non-conserved regions. Phylogenetic trees for the conserved and non-conserved regions were constructed using maximum likelihood as implemented in RAxML version 7.3.0 with the GTR + GAMMA model[63]. Standardized Robinson–Foulds distances[64] were calculated between the conserved and non-conserved trees based on 500 bootstrap replicates. Trees were considered congruent if they had no conflicting branches with bootstrap support of $>95\%$.

Phylogeny. Phylogenetic trees were constructed for the conserved OrthoMCL OG5_0126701 using the Neighbour-Joining algorithm with 1,000 bootstrap replicates as implemented in Geneious ver. 7.1.5 (http://www.geneious.com (ref. 65)) and visualized with FigTree ver. 1.4.0 (http://tree.bio.ed.ac.uk/software/figtree/).

Synteny. The OrthoMCL ortholog clusters (see above) were reformatted to represent each pair found in the cluster outside of self-matches and syntenic blocks were generated between all combinations of genomes as described in ref. 28.

Chromosome painting. Local admixture analyses using an enhanced ADMIXTURE algorithm[40] was used to assign each of the 62 strains to clusters representing these ancestral states.

Additional methodology. Detailed methods for the above sections can be found in the supplementary methods.

References

1. Baldauf, S. L. The deep roots of eukaryotes. *Science* **300,** 1703–1706 (2003).
2. Levine, N. D. *The Protozoan Phylum Apicomplexa* (CRC Press, 1988).
3. Dubey, J. P. *Toxoplasma, Hammondia, Besniotia, Sarcocystis*, and other tissue cyst-forming coccidia of man and animals. in *Parasitic Protozoa* (ed. Kreier, J. P.) 101–237 (Academic Press, 1977).
4. Berney, C. & Pawlowski, J. A molecular time-scale for eukaryote evolution recalibrated with the continuous microfossil record. *Proc. Biol. Sci.* **273,** 1867–1872 (2006).
5. Dubey, J. P. *Toxoplasmosis of Animals and Humans* 313 (CRC Press, 2010).
6. Pappas, G., Roussos, N. & Falagas, M. E. Toxoplasmosis snapshots: global status of *Toxoplasma gondii* seroprevalence and implications for pregnancy and congenital toxoplasmosis. *Int. J. Parasitol.* **39,** 1385–1394 (2009).
7. Montoya, J. G. & Liesenfeld, O. Toxoplasmosis. *Lancet* **363,** 1965–1976 (2004).
8. Weiss, L. M. & Kim, K. *Toxoplasma gondii: The Model Apicomplexan: Perspectives and Methods* 1085 (Academic Press, 2014).
9. Su, C. *et al.* Recent expansion of *Toxoplasma* through enhanced oral transmission. *Science* **299,** 414–416 (2003).
10. Sibley, L. D. & Ajioka, J. W. Population structure of *Toxoplasma gondii*: Clonal expansion driven by infrequent recombination and selective sweeps. *Annu. Rev. Microbiol.* **62,** 329–351 (2008).
11. Khan, A. *et al.* Genetic analyses of atypical *Toxoplasma gondii* strains reveals a fourth clonal lineage in North America. *Int. J. Parasitol.* **41,** 645–655 (2011).
12. Hunter, C. A. & Sibley, L. D. Modulation of innate immunity by *Toxoplasma gondii* virulence effectors. *Nat. Rev. Microbiol.* **10,** 766–778 (2012).
13. Walzer, K. A. *et al. Hammondia hammondi*, an avirulent relative of *Toxoplasma gondii*, has functional orthologs of known *T. gondii* virulence genes. *Proc. Natl Acad. Sci. USA* **110,** 7446–7451 (2013).

14. Blazejewski, T. *et al.* Systems based analysis of the *Sarcocystis neurona* genome identifies pathways that contribute to a heteroxenous life cycle. *MBio* **6,** 02445-14 (2015).
15. Reid, A. J. *et al.* Comparative genomics of the apicomplexan parasites *Toxoplasma gondii* and *Neospora caninum*: coccidia differing in host range and transmission strategy. *PLoS Pathog.* **8,** e1002567 (2012).
16. 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).
17. 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).
18. Reid, A. J. *et al.* Genomic analysis of the causative agents of coccidiosis in domestic chickens. *Genome Res.* **24,** 1676–1685 (2014).
19. Chen, F., Mackey, A. J., Stoeckert, Jr C. J. & Roos, D. S. OrthoMCL-DB: Querying a comprehensive multi-species collection of ortholog groups. *Nucleic Acids Res* **34,** D363–D368 (2006).
20. Sonnhamer, E. L. L., Eddy, S. R. & Durbin, R. Pfam: a comprehensive database of protein families based on seed alignments. *Proteins* **3,** 405–420 (1997).
21. Miranda-Saavedra, D., Gabaldon, T., Barton, G. J., Langsley, G. & Doerig, C. The kinomes of apicomplexan parasites. *Microbes Infect.* **14,** 796–810 (2012).
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-integrase DNA binding domains. *Nucleic Acids Res.* **33,** 3994–4006 (2005).
24. Wasmuth, J. D. *et al.* Integrated bioinformatic and targeted deletion 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 s48/45 domain. *Proc. Natl Acad. Sci. USA* **109,** 6692–6697 (2012).
26. Tonkin, M. L. *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. DeBarry, J. D. & Kissinger, J. C. Jumbled genomes: missing Apicomplexan synteny. *Mol. Biol. Evol.* **28,** 2855–2871 (2011).
29. Behnke, M. S. *et al.* Virulence differences in *Toxoplasma* mediated by amplification of a family of polymorphic pseudokinases. *Proc. Natl Acad. Sci. USA* **108,** 9631–9636 (2011).
30. Taylor, S. *et al.* A secreted serine-threonine kinase determines virulence in the eukaryotic pathogen *Toxoplasma gondii*. *Science* **314,** 1776–1780 (2006).
31. Selmecki, A. M. *et al.* Polyploidy 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 Toxoplasmatinae species and closely related strains of *Toxoplasma gondii*. *mBio* **5,** e01003–e01013 (2014).
33. Cowper, B., 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* subtelomeric-like regions: identification of a long-range compositional bias that is also associated with gene-poor regions. *BMC Genomics* **15,** 21 (2014).
35. Scherf, A., Lopez-Rubio, J. J. & Riviere, L. Antigenic variation in *Plasmodium falciparum*. *Annu. Rev. Microbiol.* **62,** 445–470 (2008).
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 transcontinental 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 ancestry in unrelated individuals. *Genome Res* **19,** 1655–1664 (2009).
41. Zhou, 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).
43. Lawson, D. J., Hellenthal, G., Myers, S. & Falush, D. Inference of population structure using dense haplotype data. *PLoS Genet.* **8,** e1002453 (2012).
44. Etheridge, R. D., Alagan, A., Tang, K., Turk, B. E. & Sibley, L. D. ROP18 and ROP17 kinase complexes synergize to control acute virulence of Toxoplasma in the mouse. *Cell Host Microbe* **15,** 537–550 (2014).
45. Craver, M. P. & Knoll, L. J. Increased efficiency of homologous recombination in *Toxoplasma gondii* dense granule protein 3 demonstrates that GRA3 is not necessary in cell culture but does contribute to virulence. *Mol. Biochem. Parasitol.* **153,** 149–157 (2007).
46. Radke, J. R. *et al.* Identification of a sporozoite-specific member of the *Toxoplasma* SAG superfamily via genetic complementation. *Mol. Microbiol.* **52,** 93–105 (2004).
47. Dzierszinski, F., Mortuaire, M., Cesbron-Delauw, M. F. & Tomavo, S. Targeted disruption of the glycosylphosphatidylinositol-anchored surface antigen SAG3 gene in *Toxoplasma gondii* decreases host cell adhesion and drastically reduces virulence in mice. *Mol. Microbiol.* **37,** 574–582 (2000).
48. Boyle, J. P. *et al.* 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 (2006).
49. Wendte, J. M. *et al.* Self-mating in the definitive *host potentiates* clonal outbreaks of the apicomplexan parasites *Sarcocystis neurona* and *Toxoplasma gondii*. *PLoS Genet.* **6,** e1001261 (2010).
50. Gardner, M. J. *et al.* Genome sequence of the human malaria parasite *Plasmodium falciparum*. *Nature* **419,** 498–511 (2002).
51. Carlton, J. M. *et al.* Comparative genomics of the neglected human malaria parasite *Plasmodium vivax*. *Nature* **455,** 757–763 (2008).
52. Berriman, M. *et al.* The genome of the African trypanosome *Trypanosoma brucei*. *Science* **309,** 416–422 (2005).
53. Baxter, L. *et al.* Signatures of adaptation to obligate biotrophy in the *Hyaloperonospora arabidopsidis* genome. *Science* **330,** 1549–1551 (2010).
54. Miller, J. R., Koren, S. & Sutton, G. Assembly algorithms for next-generation sequencing data. *Genomics* **95,** 315–327 (2010).
55. Delcher, A. L., Phillippy, A., Carlton, J. & Salzberg, S. L. Fast algorithms for large-scale genome alignment and comparison. *Nucleic Acids Res.* **30,** 2478–2483 (2002).
56. Marchler-Bauer, A. *et al.* CDD: conserved domains and protein three-dimensional structure. *Nucleic Acids Res.* **41,** D348–D352 (2013).
57. Eddy, S. R. Accelerated Profile HMM Searches. *PLoS Comput. Biol.* **7,** e1002195 (2011).
58. Kall, L., Krogh, A. & Sonnhammer, E. L. A combined transmembrane topology and signal peptide prediction method. *J. Mol. Biol.* **338,** 1027–1036 (2004).
59. Bailey, T. L. *et al.* MEME SUITE: tools for motif discovery and searching. *Nucleic. Acids Res.* **37,** W202–W208 (2009).
60. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9,** 357–359 (2012).
61. Li, L., Stoeckert, Jr C. J. & Roos, D. S. OrthoMCL: identification of ortholog groups for eukaryotic genomes. *Genome Res.* **13,** 2178–2189 (2003).
62. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics.* **25,** 2078–2079 (2009).
63. Stamatakis, A. RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics.* **22,** 2688–2690 (2006).
64. Robinson, D. R. & Foulds, L. R. Comparison of phylogenetic trees. *Math. Biosci.* **53,** 131–147 (1981).
65. Kearse, M. *et al.* Geneious Basic: an integrated and extendable desktop software platform for the organization and analysis of sequence data. *Bioinformatics.* **28,** 1647–1649 (2012).
66. Buchholz, K. R. *et al.* Identification of tissue cyst wall components by transcriptome analysis of in vivo and in vitro *Toxoplasma gondii* bradyzoites. *Eukaryot. Cell* **10,** 1637–1647 (2011).
67. Fritz, H. M. *et al.* Transcriptomic analysis of toxoplasma development reveals many novel functions and structures specific to sporozoites and oocysts. *PLoS ONE* **7,** e29998 (2012).

Acknowledgements

Sequencing, assembly and annotation for the Tg and Hh genomes reported here were supported by the National Institute of Allergy and Infectious Diseases (NIAID), National Institutes of Health (NIH), Department of Health and Human Services under contract number HHSN272200900007C. Additional analyses were provided by community-based efforts and included support from NIH grants AI059176 and AI036629 (L.D.S.) and the Canadian Institutes for Health Research grant MOP 84556 (J.P.). Data for the Sn annotation was generously provided by M.E.G. and J.P. prior to publication and was supported by the Intramural Research Program of NIAID at the NIH (M.E.G.). Additional computing resources were provided by the SciNet HPC Consortium. We acknowledge ToxoDB (http://toxodb.org/) for providing a publically available repository for all genomic data described here and for assistance in coordinating this study. OrthoMCL and ToxoDB are supported in part with funds from the NIAID, NIH, and Department of Health and Human Services under contract HHSN272201400030C (D.S.R. & J.C.K.).

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., B.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., J.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., D.K.H., K.K., B.M.R., M.W.W., J.P., J.C.K., J.P.J.S., 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), TgCtBr5 (GCA_000259835.1), P89 (GCA_000224885.2), VAND (GCA_000224845.2), COUG (GCA_000338675.1), ARI (GCA_000250965.1), TgCtPRC2 (GCA_000256725.1), GAB2-2007-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-PaperProvisionalGenomes/): 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), TgCkCr1 (SRX160807), SOU (SRX160119), TgCkGy2 (SRX099785), CASTELLS (SRX099789), TgCatBr1 (SRX099790), TgCatBr18 (SRX099794), TgCatBr34 (SRX160051), BRC TgH 18002 GUY-KOE (SRX099796), BRC TgH 18003 GUY-MAT (SRX099783), GUY-2003-MEL (SRX160131), GUY-2004-ABE (SRX160132), TgRsCr1 (SRX160143), TgCkBr141 (SRX160124), TgCatBr25 (SRX160134), TgCatBr10 (SRX099791), TgCatBr64 (SRX160743), BRC TgH 18001 GUY-DOS (SRX099782), TgCkCr10 (SRX099784), B41 (SRX099774), RAY (SRX099793), TgCatBr44 (SRX160141), GAB1-2007-GAL-DOM10 (SRX159841, SRX159839), GAB3-2007-GAL-DOM9 (SRX160125), GAB5-2007-GAL-DOM6 (SRX159842, SRX159840), TgCatPRC3 (SRX09979), BRC TgH 18009 (SRX171132), BRC TgH 18021 (SRX160041), BRC TgH 20005 (SRX157499, SRX157465), BRC TgH 21016 (SRX099775), GUY-2004-JAG1 (SRX099776), B73 (SRX159844, SRX159843), PRU (SRX099792), M7741 (SRX159890, SRX159849), ROD (SRX160129, SRX160128), TgShUS28 (SRX160130), BOF (SRX099774), GAB5-2007-GAL-DOM1 (SRX160069), TgCatBr26 (SRX099780), TgCatBr72 (SRX160049), TgDogCo17 (SRX099787), TgCatBr15 (SRX099779), G662M (SRX160052), GAB3-2007-GAL-DOM2 (SRX160123), TgCatBr3 (SRX160142) and TgH26044 (SRX160050).

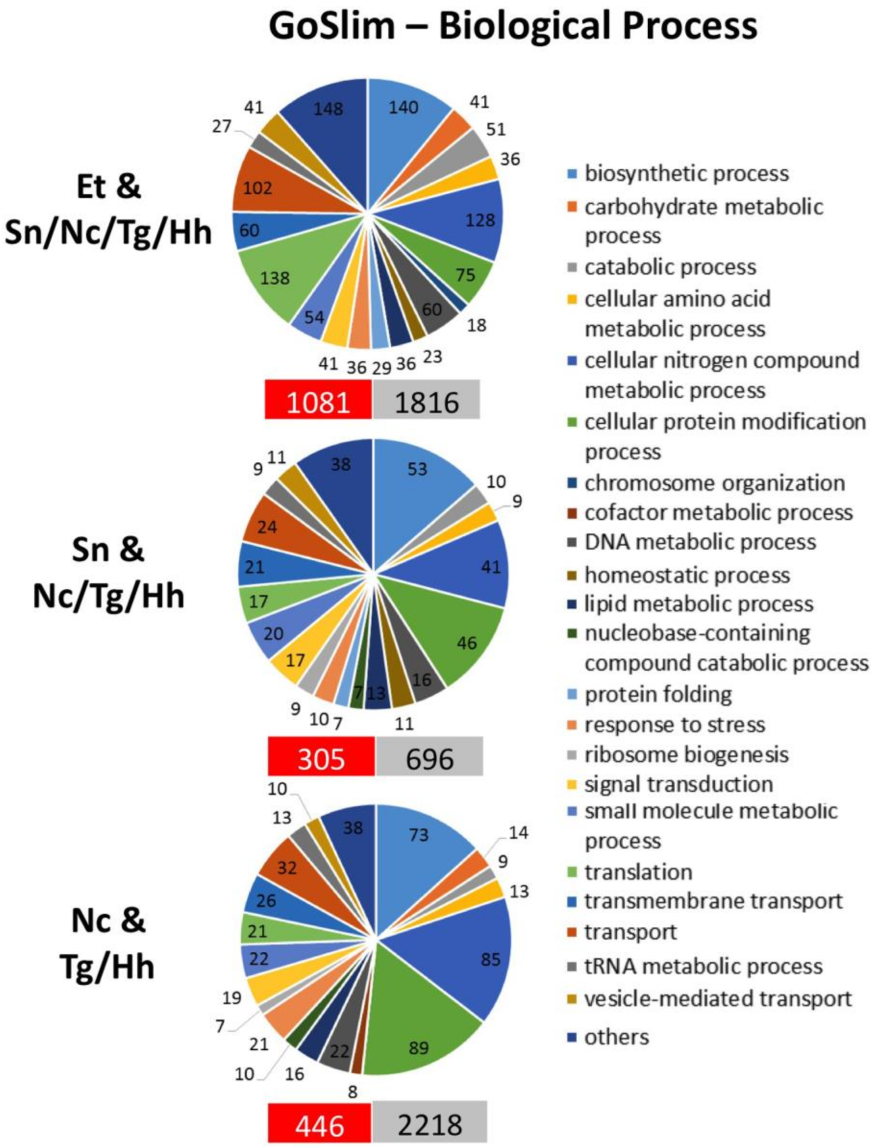
Supplementary Information accompanies this paper at http://www.nature.com/naturecommunications

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

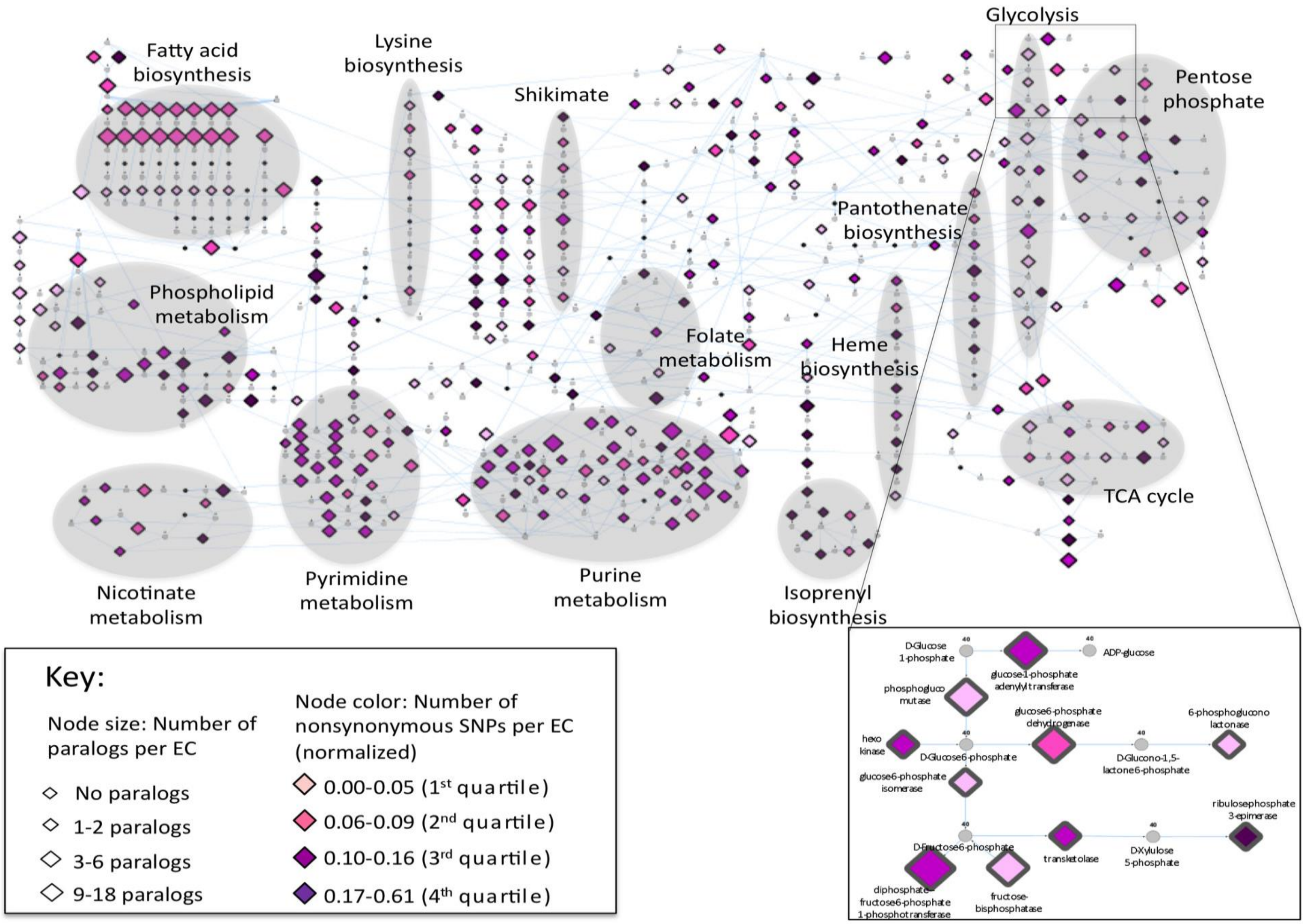
Reprints and permission information is available online at http://npg.nature.com/reprintsandpermissions/

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).

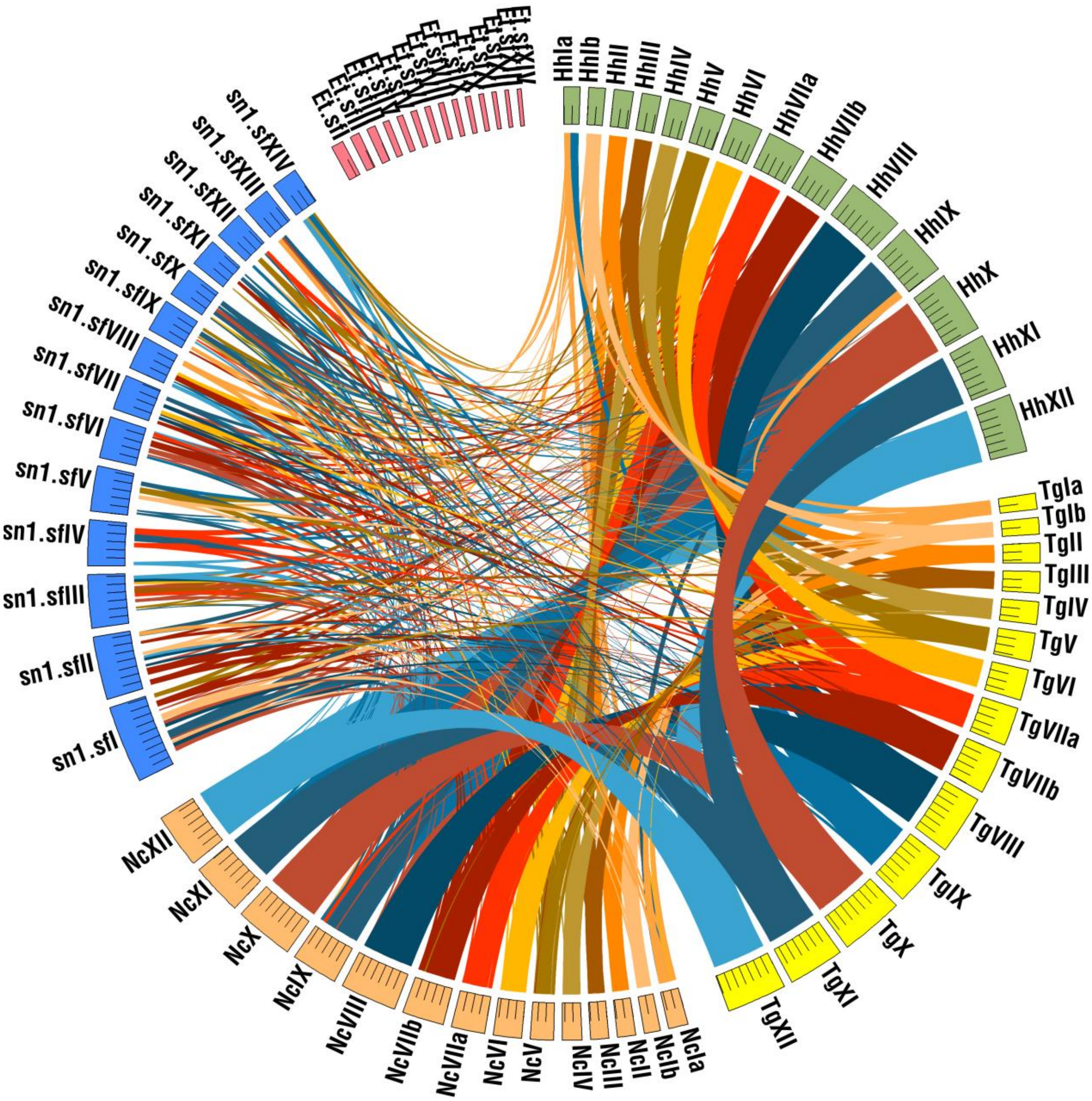
Supplementary Figures, Tables, Methods, and References



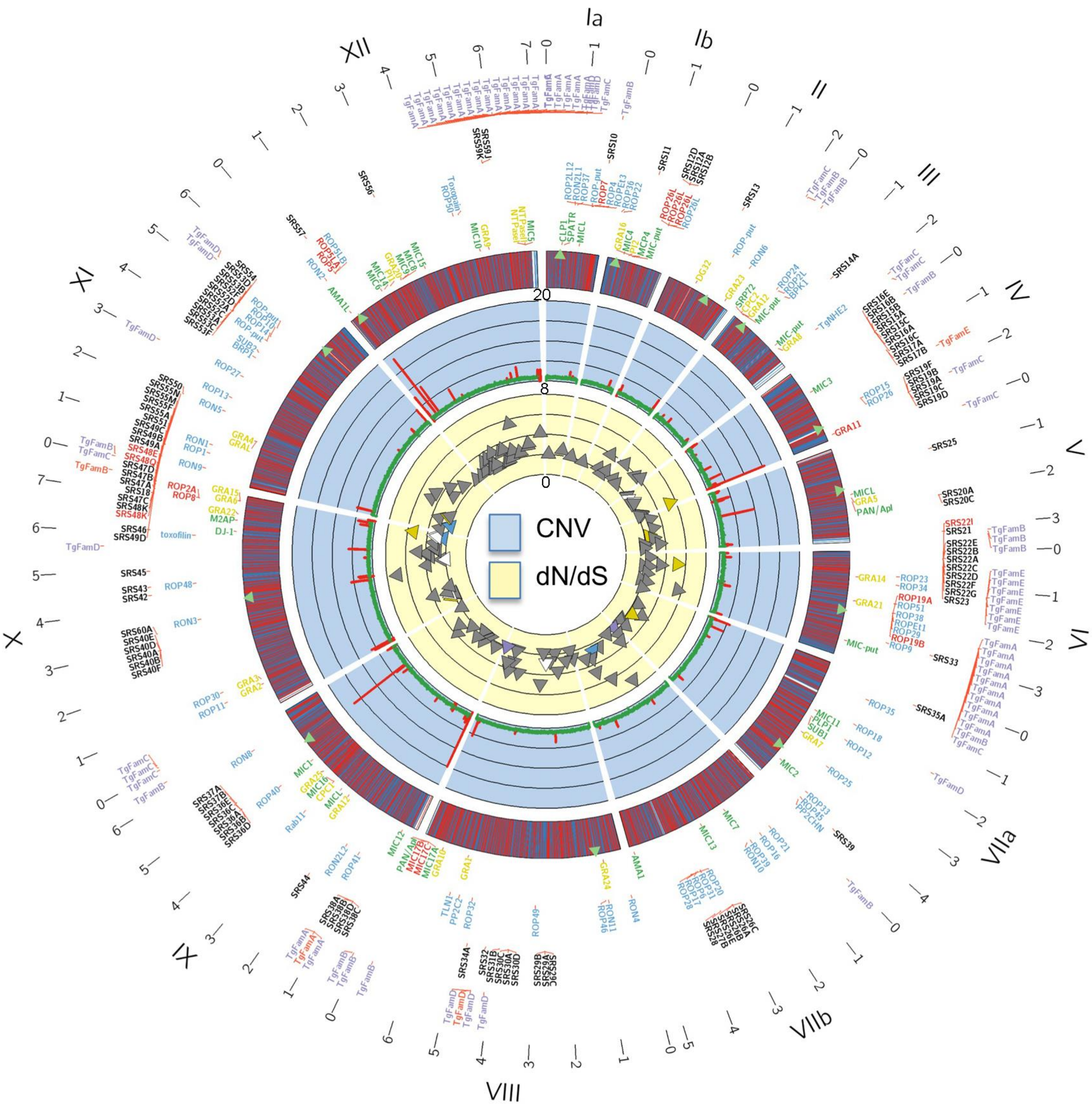
Supplemental Figure 1 Pie-charts representing the top 20 GO Slim annotations (Biological Process) of clusters of orthologous groups. At each branch, OrthoMCL groups present in the outer species and at least one of the internal taxa are included. Numbers in grey indicate orthogroups without pfam domains, numbers in red indicate those with pfam annotations. Et – *E. tenella*, Sn - *S. neurona*, Nc - *N. caninum*, Tg - *T. gondii*, Hh - *H. hammondi*. As expected, all 5 coccidians share orthologous groups corresponding to many key biological processes. However, it is noteworthy to see that the group of tissue-cyst coccidians (Sn ,Nc, Tg, Hh) also share 1,001 orthogroups not present in Et. In particular, they are enriched for processes involved in protein modification. This trend is also observed in the orthogroups shared by the closely related set of Nc, Tg, and Hh (2,664 orthogroups in common), suggesting a key role for processes involving protein modification during their evolution.


Glycolysis
Fatty acid biosynthesis
Lysine biosynthesis
Shikimate
Pentose phosphate
Pantothenate biosynthesis
Phospholipid metabolism
Folate metabolism
Heme biosynthesis
TCA cycle
Nicotinate metabolism
Pyrimidine metabolism
Purine metabolism
Isoprenyl biosynthesis
Key:
Node size: Number of paralogs per EC
◇ No paralogs
◇ 1-2 paralogs
◇ 3-6 paralogs
◇ 9-18 paralogs
Node color: Number of nonsynonymous SNPs per EC (normalized)
◇ 0.00-0.05 (1st quartile)
◆ 0.06-0.09 (2nd quartile)
◆ 0.10-0.16 (3rd quartile)
◆ 0.17-0.61 (4th quartile)
D-Glucose 1-phosphate
40
ADP-glucose
glucose-1-phosphate adenylyltransferase
phosphogluco mutase
glucose6-phosphate dehydrogenase
6-phosphoglucono lactonase
hexo kinase
D-Glucose6-phosphate
D-Glucono-1,5-lactone6-phosphate
glucose6-phosphate isomerase
ribulosephosphate 3-epimerase
D-Fructose6-phosphate
transketolase
D-Xylulose 5-phosphate
diphosphate--fructose6-phosphate 1-phosphotransferase
fructose-bisphosphatase

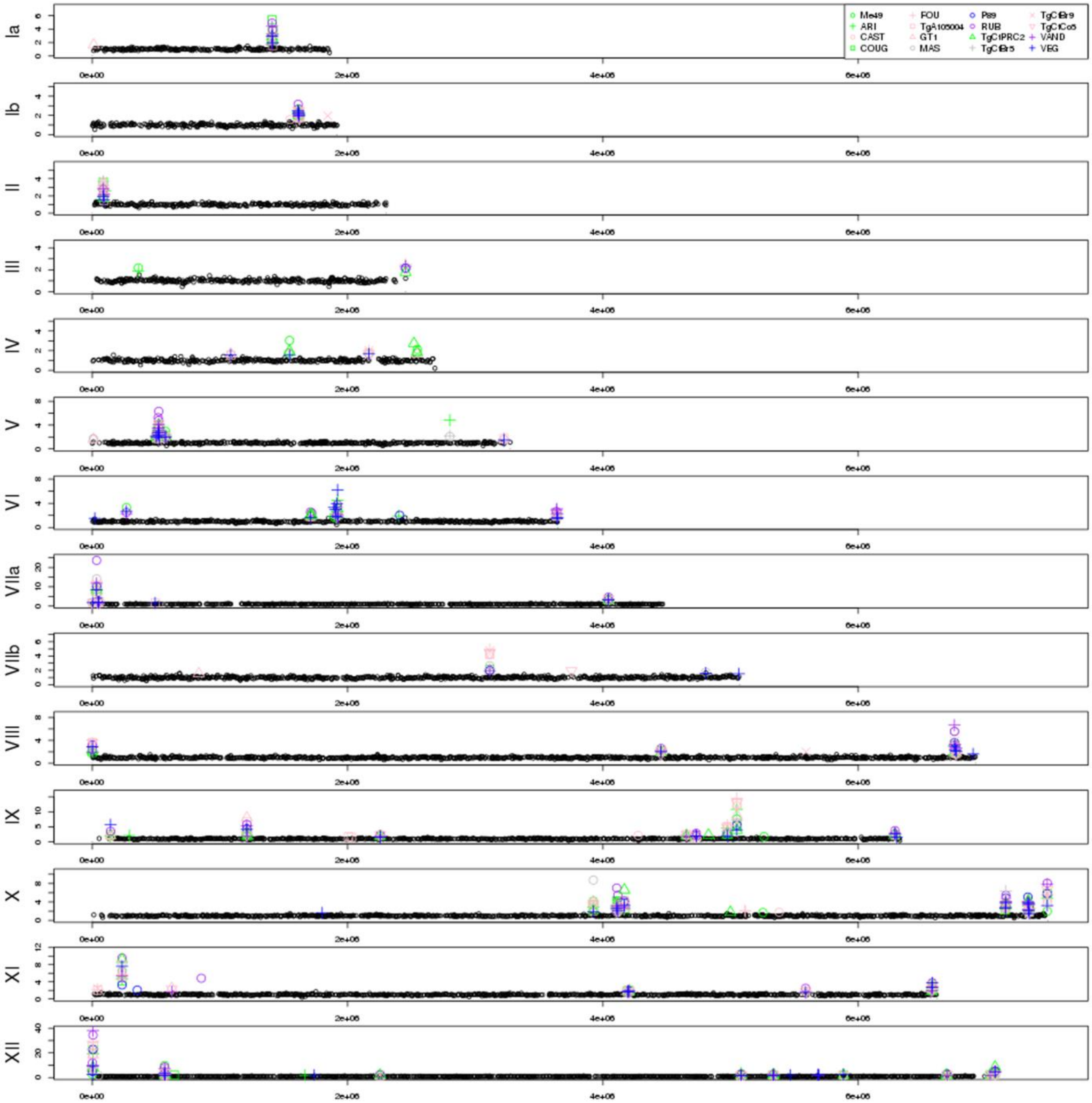
Supplemental Figure 2 Metabolic network of *T. gondii* (iCS382) annotated based on non-synonymous SNPs and number of paralogs. The metabolic network reconstruction of *T. gondii* (iCS382) is shown with nodes representing enzymes (diamonds) and metabolites (circles), connected by edges. Groups of enzymes constituting individual pathways (according to KEGG) are enclosed in grey circles. The number of paralogs for each EC number is represented by node size. The frequency of non-synonymous SNPs (including nonsense SNPs) for each EC number is represented by node color. These values are normalized for the length of the protein and number of paralogs. The region corresponding to the beginning of the energy metabolism pathways glycolysis and pentose phosphate pathway is enlarged and shown as an inset.



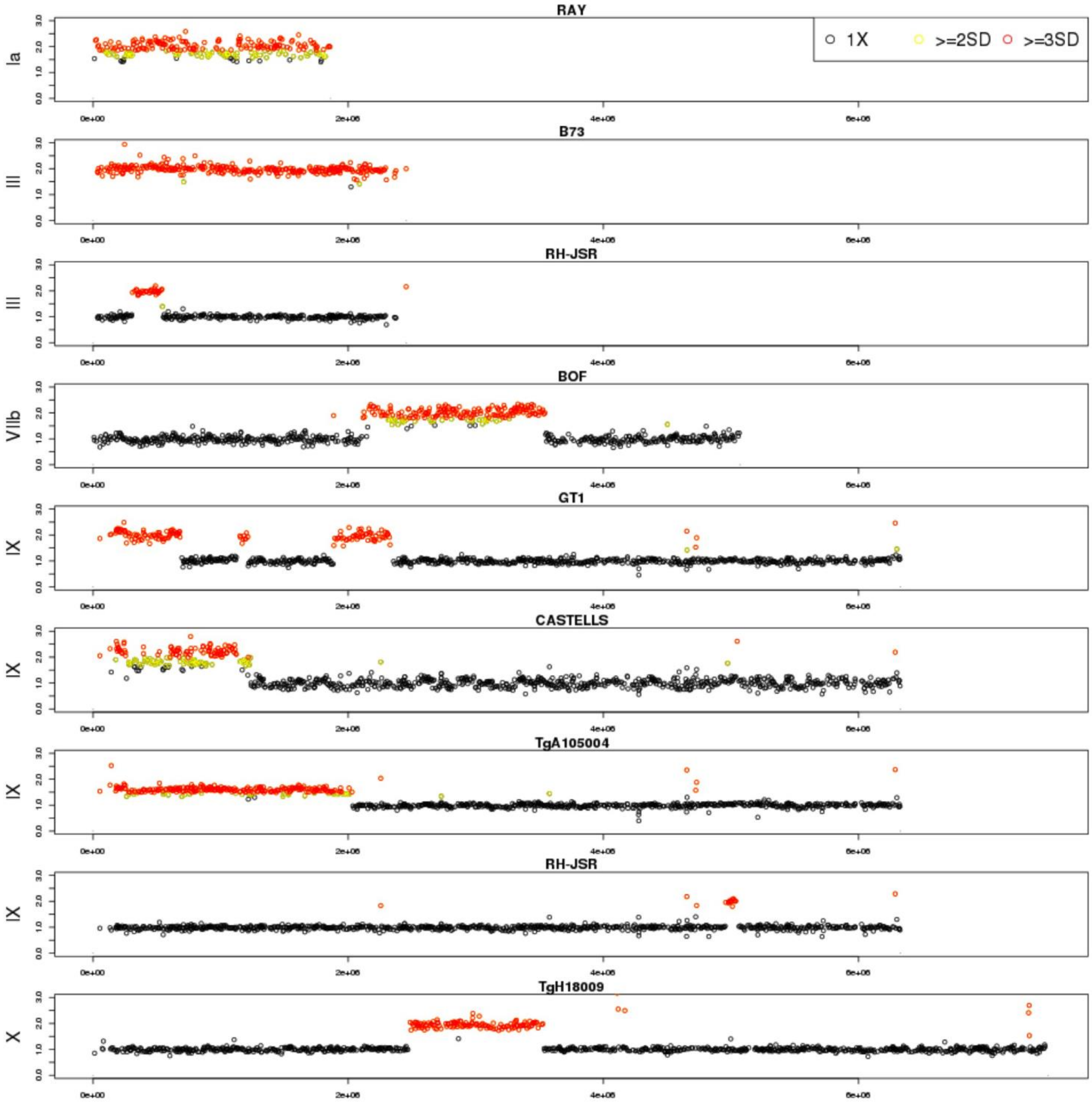
Supplemental Figure 3 Levels of synteny (relaxed criteria) among the coccidia. The large outer circle represents the annotated chromosomes or scaffolds of each coccidian species. For *T. gondii, H. hammondi* and *N. caninum* all assembled chromosomes (n=14) are plotted. For *S. neurona* and *E. tenella* the largest 14 scaffolds are plotted. Each chromosome/scaffold is labeled with the genus-species abbreviation followed by the chromosome/scaffold number. Tick marks on the chromosome/scaffold represent 1 Mb. The colored bands linking chromosome/scaffold pairs represent syntenic blocks (minimum of 3 genes) shared by the species 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 orthologous clusters, where the cluster contained at least two genes, i.e. the gene is not present in all species. Data summarized in **Supplemental Table 2**.



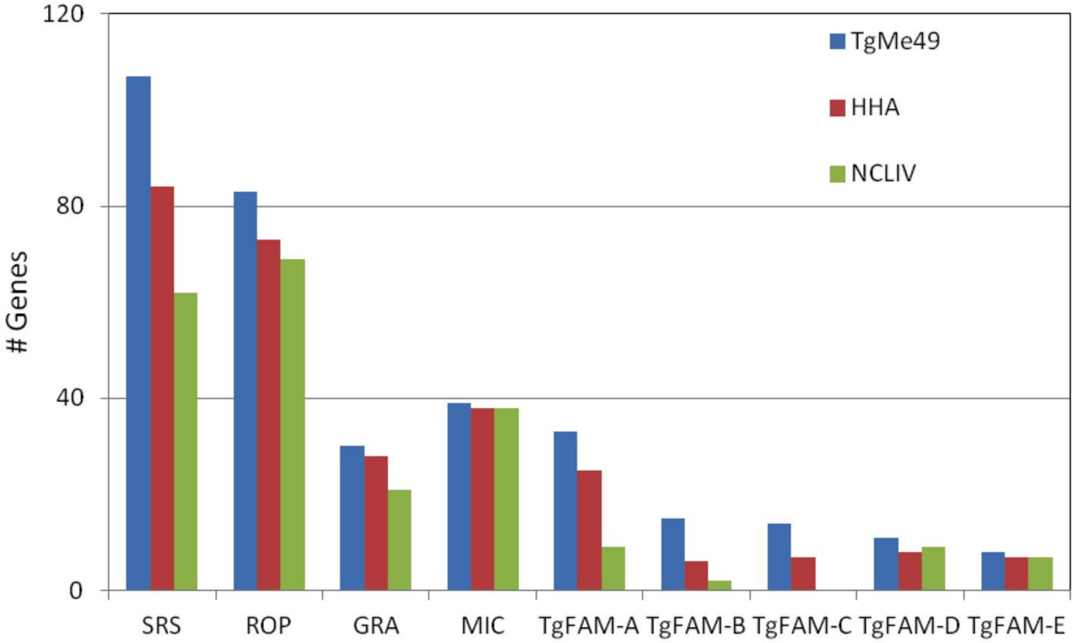
Supplemental Figure 4 Circos representational plot of Toxoplasma genome. The innermost yellow track is a scatter plot of genes with dN/dS ratio ≥ 2 (136 genes): y-axis 0 - 8, colors indicate SPD family. The track with blue background is a histogram of regions with CNV in ME49 using a rolling window of 2,500 bp: y-axis 0 - 20, 1X coverage (green), regions with CNV (red). The third circular track from the center plots the genes on 14 chromosomes: top strand gene (blue), bottom strand gene (red), centromere locations (green triangles). Outside of these circular tracks the locations of SPD genes are indicated in text. These categories ordered from the inside out are: *GRA* (yellow – 28 locations with 31 genes), *MIC* (green – 33 locations with 42 genes), *ROP* (blue – 71 locations with 104 genes), *SRS* (black – 47 locations with 115 genes), *TgFAM* (purple – 42 locations with 83 genes). In addition, genes with CNV are shown in red. Chromosome numbers (Roman) are shown on the outer ring, with size demarcations indicated by the Arabic numbers and tick marks.



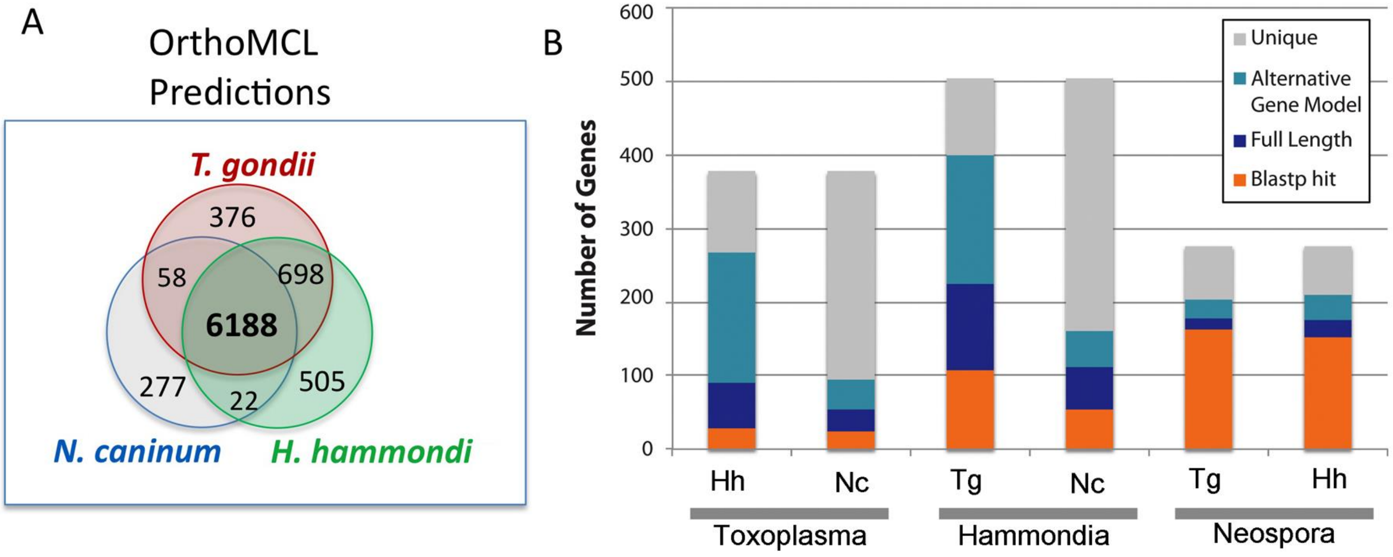
Supplemental Figure 5 Distribution of CNV in the 16 Toxoplasma reference genomes. Chromosome level plots of genes with CNV in the 16 Toxoplasma reference genomes. ME49 genes with no CNV (1X) are plotted by chromosomal location with black circles. Genes with CNV in each of the 16 reference genomes are plotted by chromosomal location with color based on Clade membership: Clade A (pink), Clade B (gray), Clade C (blue), Clade D (green), Clade E (yellow), Clade F (purple).



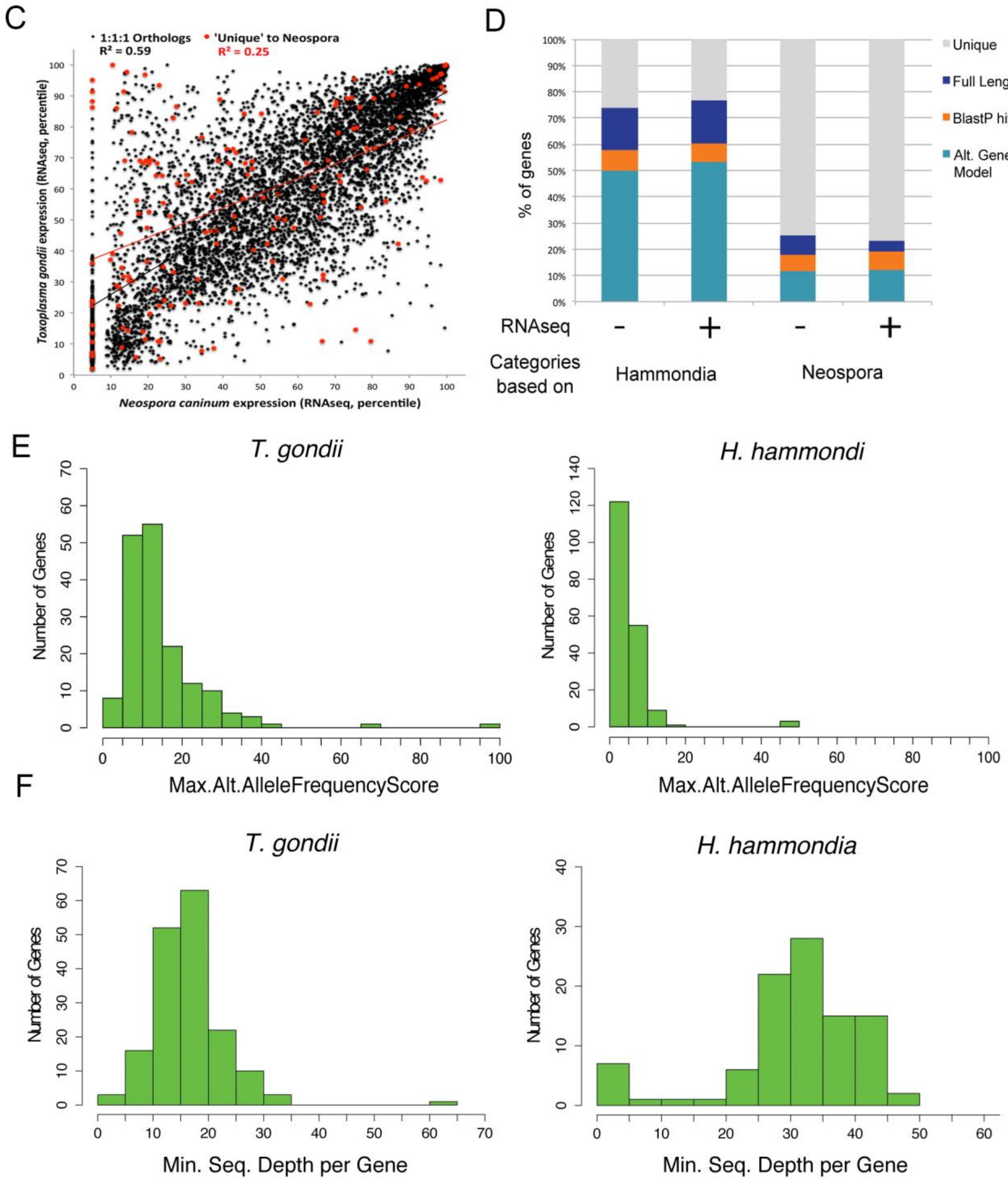
Supplemental Figure 6 Strains with aneuploidy or large duplicated regions. Select chromosomes are shown for Individual strains where 10 or more genes with CNV occur continuously. 1X genes (black), genes with > 2 X standard deviation (SD) CNV (yellow - borderline CNV), genes with > 3 X standard deviation (SD) CNV (red).



Supplemental Figure 7 Distribution of orthologous genes by species. Total number of genes for each category found in each of *T. gondii* (ME49 strain), *H. hammondi* (HHa) or *N. caninum* (NCLIV). Based on OrthoMCL clustering.



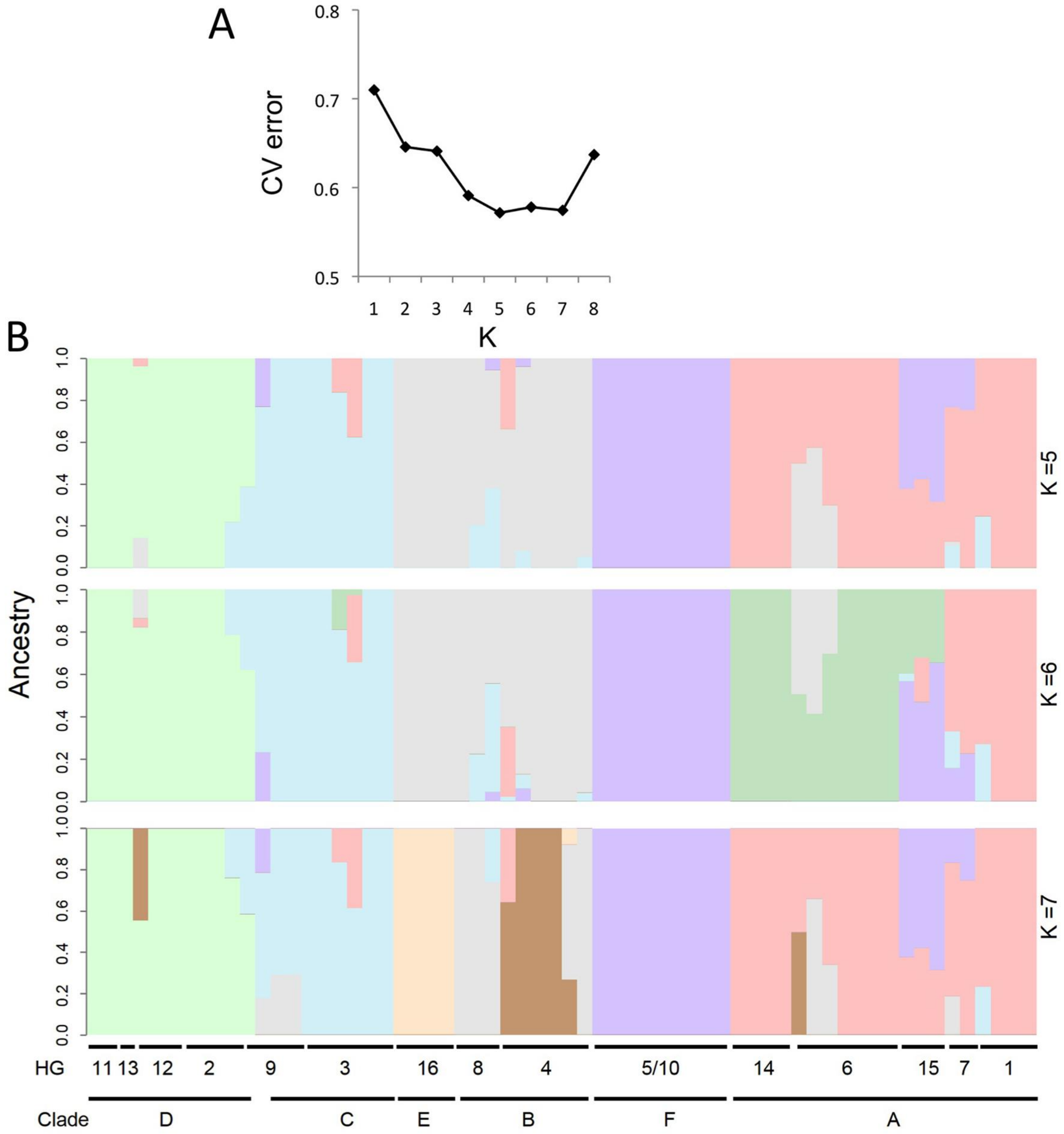
Supplemental Figure 8 Conserved genes among closely related tissue cyst forming coccidia. A) Number of shared and species-specific (unique) clusters of orthologous genes as defined by OrthoMCL. The number of clusters of orthologous genes shown for *T. gondii* represents those found in all 16 references strains. B) Analysis of species-specific genes as predicted by OrthoMCL. Orange bars (Blastp hit) represent species-specific genes that identify orthologs in the proteomes of the other two species with a blastp e-value ≤ $1x10^{-10}$ and 50% coverage of the shortest sequence. Blue bars (Full Length) depict species-specific genes that map as syntenic hits to the genomes of the other two species with at least 70% coverage using either GMAP or tblastn (e-value ≤ $1x10^{-5}$). These later two categories are likely distant orthologs that are classified as separate groups by OrthoMCL. Turquoise bars (Alternative Gene Model) represent the same as blue bars but for species-specific genes that map with less than 70% coverage or contain non-sense mutations or frame-shifts in the target genome. Gray bars (Unique) depict species-specific genes that do not match the proteomes or genomes of the other two species. Tg, *T. gondii*; Hh, *H. hammondi*; Nc, *N. caninum*.



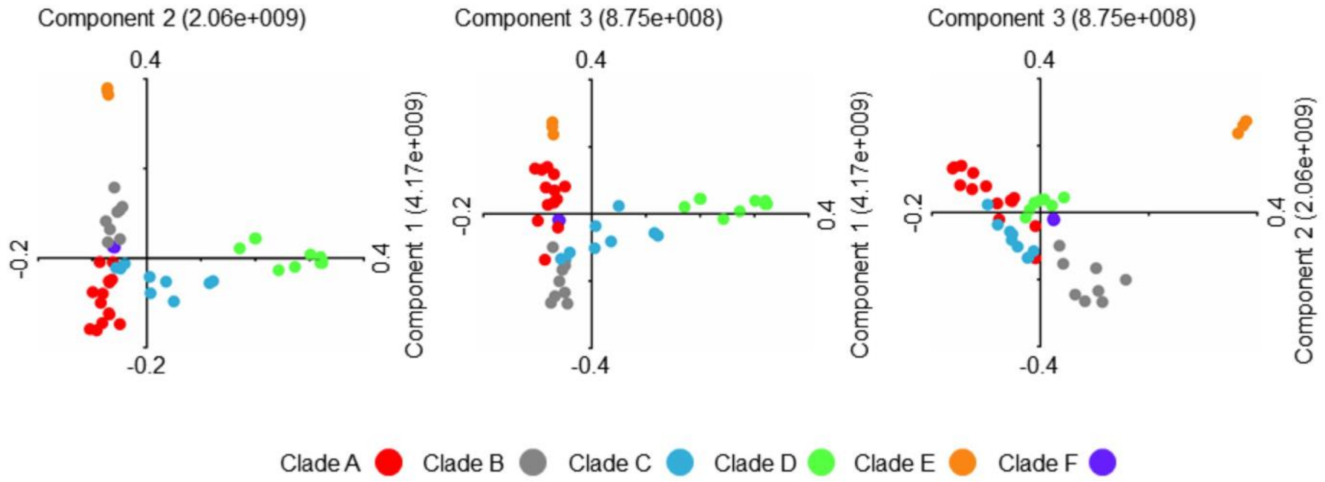
Supplemental Figure 8 Conserved genes among closely related tissue cyst forming coccidian continued.

C) Comparative gene expression for orthologous and predicted species-specific genes in *N. caninum* and *T. gondii*. For genes were there is a 1:1:1 orthologous correspondence between *N. caninum*, *T. gondii* and *H. hammondi* there was a good correlation of RNA-expression levels, based on RNA-seq data, between *N. caninum* and *T. gondii* (black dots). However for genes that were originally flagged as "unique" by OrthoMCL but for which there are annotated genes in both *N. caninum* and *T. gondii*,

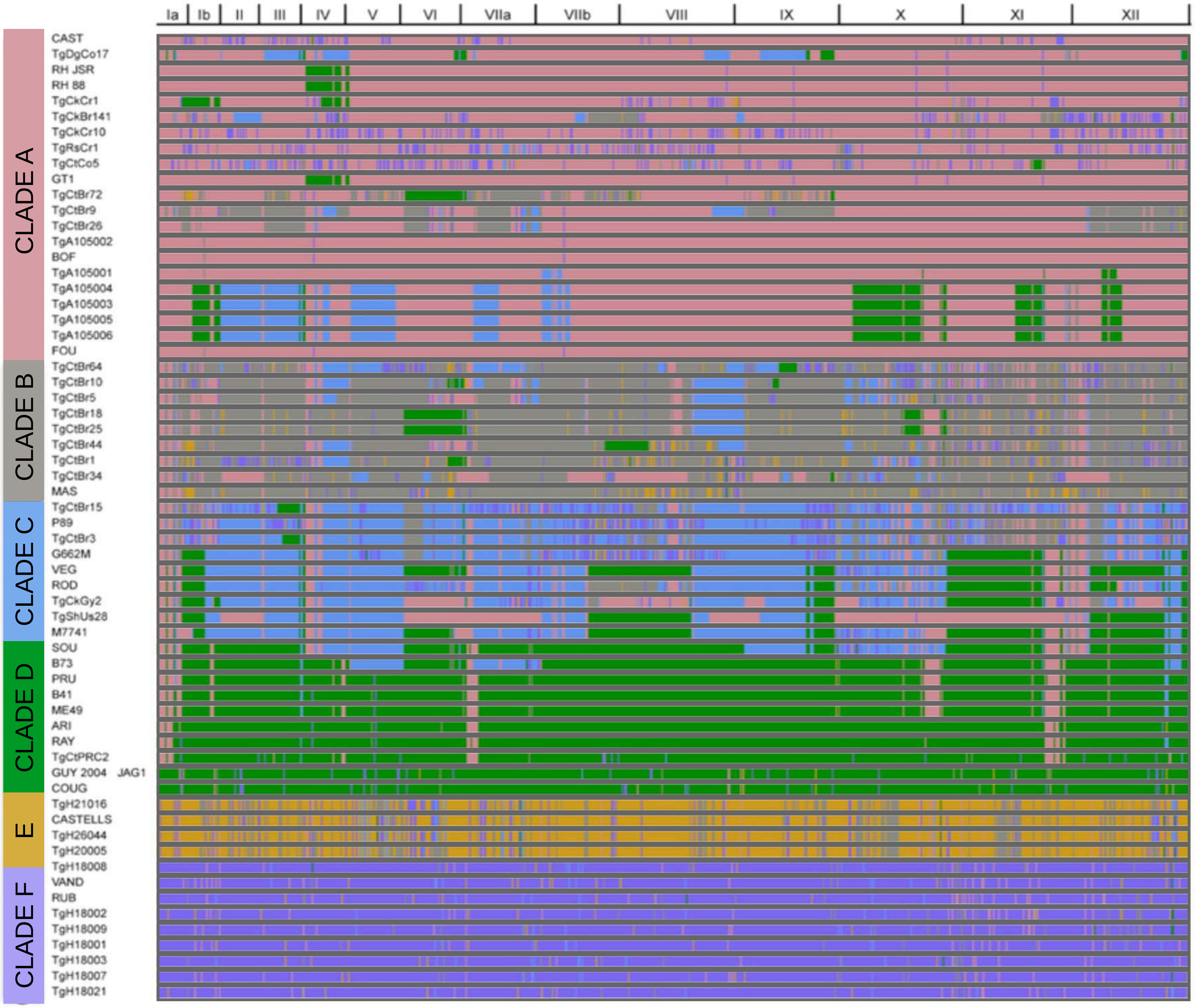
the transcript levels did not correlate well (red dots). This suggests that these genes encode functionally different products and are not the result of assembly errors or inaccurate gene models.
D) RNA-seq analysis monitoring gene expression of putative species-specific genes in *T. gondii*. Plot shows the fraction of *T. gondii* genes from each of the four categories depicted in B that are supported (+) or not (-) by RNA-seq data. Potential *T. gondii* specific genes are defined by comparison to *H. hammondi* (left) and *N. caninum* (right). The expectation is that if unique and alternative gene models predicted in *T. gondii* were artifactual, they would be enriched in regions that lack RNA-seq data, compared to genes where there was evidence for orthologues being present (Full Length and BLASTP). Comparison of the genes in these different categories in regions with (+) and without (-) RNA-seq data showed no significant differences for any of the four gene categories ($P > 0.4$, Fisher's Exact Test). This result suggests that most unique genes and alternative gene models are not likely the result of artifactual gene predictions.
E) To assess if the high number of alternative gene models between *T. gondii* and *H. hammondi* was the product of genome sequencing errors we mapped genome sequencing reads to their respective *T. gondii* ME49 strain and *H. hammondi* genomes and calculated the maximum alternative allele frequency (MAAF) score within regions of putative species-specific genes. MAAF scores were defined as the maximum alternative allele frequency detected for SNPs and INDELs within a defined genomic region (see methods). It is expected that regions containing sequencing errors will present MAAF scores of ≥ 40%. This analysis revealed that only a small fraction (< 10 regions) of the loci containing potential alternative gene models in *T. gondii* (left) and *H. hammondi* (right) have MAAF scores ≥ 40%. This result suggests that frame-shifts and missense mutations found in alternative gene model loci are not the result of low sequencing quality in these regions.
F) Analysis of sequence coverage. More than 90% of the loci hit by alternative gene models in *T. gondii* and *H. hammondi* have a minimum sequencing depth higher than 5X. Left, *T. gondii* genomic regions similar to *H. hammondi* alternative gene model genes; right, *H. hammondi* genomic regions similar to *T. gondii* alternative gene model genes. These results also support the conclusion that alternative gene models are not due to poor sequencing quality.



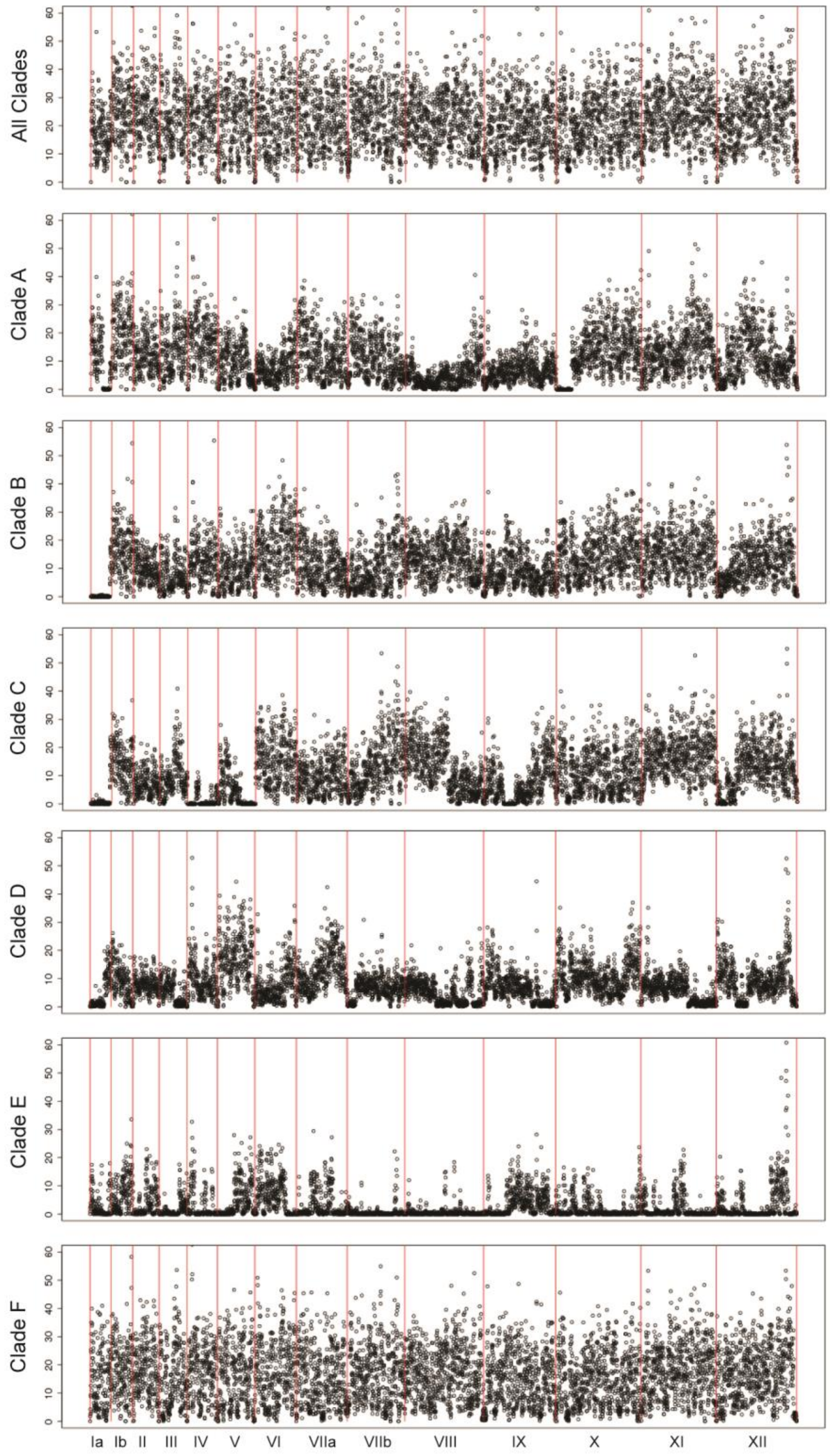
Supplemental Figure 9 Population structure inferred by Admixture analysis.
A) Cross validation (CV) plot was drawn for the whole genome SNPs data of *T. gondii*. Plot displays the CV error versus K values. CV errors for this dataset suggest K = 6 is the best fit. B) Admixture clustering analysis of *T. gondii* closely resembles the population clustering inferred by Neighbor-net analysis. Each individual in the plot is represented by a vertical stacked column of proportional genetic components of shared ancestry for K = 5 through K = 7. Each color represents each ancestral population. Previously designated haplogroups are indicated as HG.



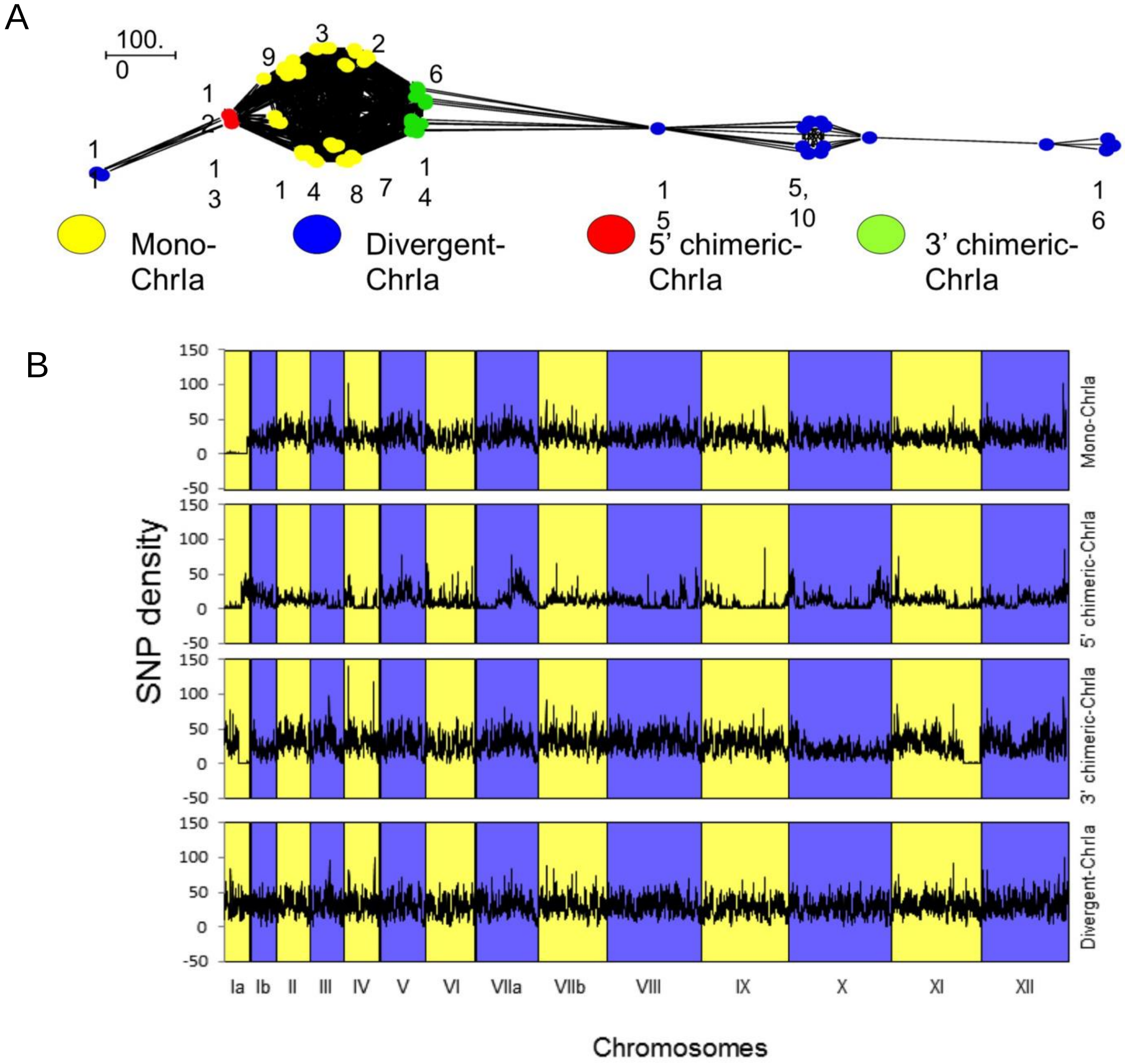
Supplemental Figure 10 Population genetic structure of *T. gondii* defined by principal component analysis. Color scheme of each clade of *T. gondii* follows **Fig. 3A**.



Supplemental Figure 11 Chromosome painting of 62 *Toxoplasma gondii* strains with local admixture analyses. Local admixture analyses were conducted on SNP blocks of size 1,000 bp on each of the 14 chromosomes. For each SNP block, local admixture assigned strains to a particular ancestral population. Sequences with the same color have high similarity, although this is not meant to imply origin. This analysis also revealed patterns of local admixture that suggest the occurrence of genetic crosses among strains of different clades, likely favored by their geographic proximity. For example, clade A strains TgCkBr141, TgCkCr10, TgRsCr1 and TgCtCo5 harbor a high number of small haploblocks that are shared with clade F strains (see purple bands for TgCkBr141, TgCkCr10, TgRsCr1 and TgCtCo5 strains and pink bands for clade F strains. Interestingly, all clade F strains were isolated from French Guiana while the aforementioned clade A strains were isolated from the nearby locations of Pará State, north of Brazil (TgCkBr141), Costa Rica (TgCkCr10 and TgRsCr1) and Colombia (TgCtCo5) (**Supplementary Data 1**). A similar high proportion of shared haploblocks occurs among several Brazilians strains from clades A (TgCtBr26, TgCtBr9 and TgCtBr72), B, and C (TgCtBr15, P89, and TgCtBr3)(see gray bands).

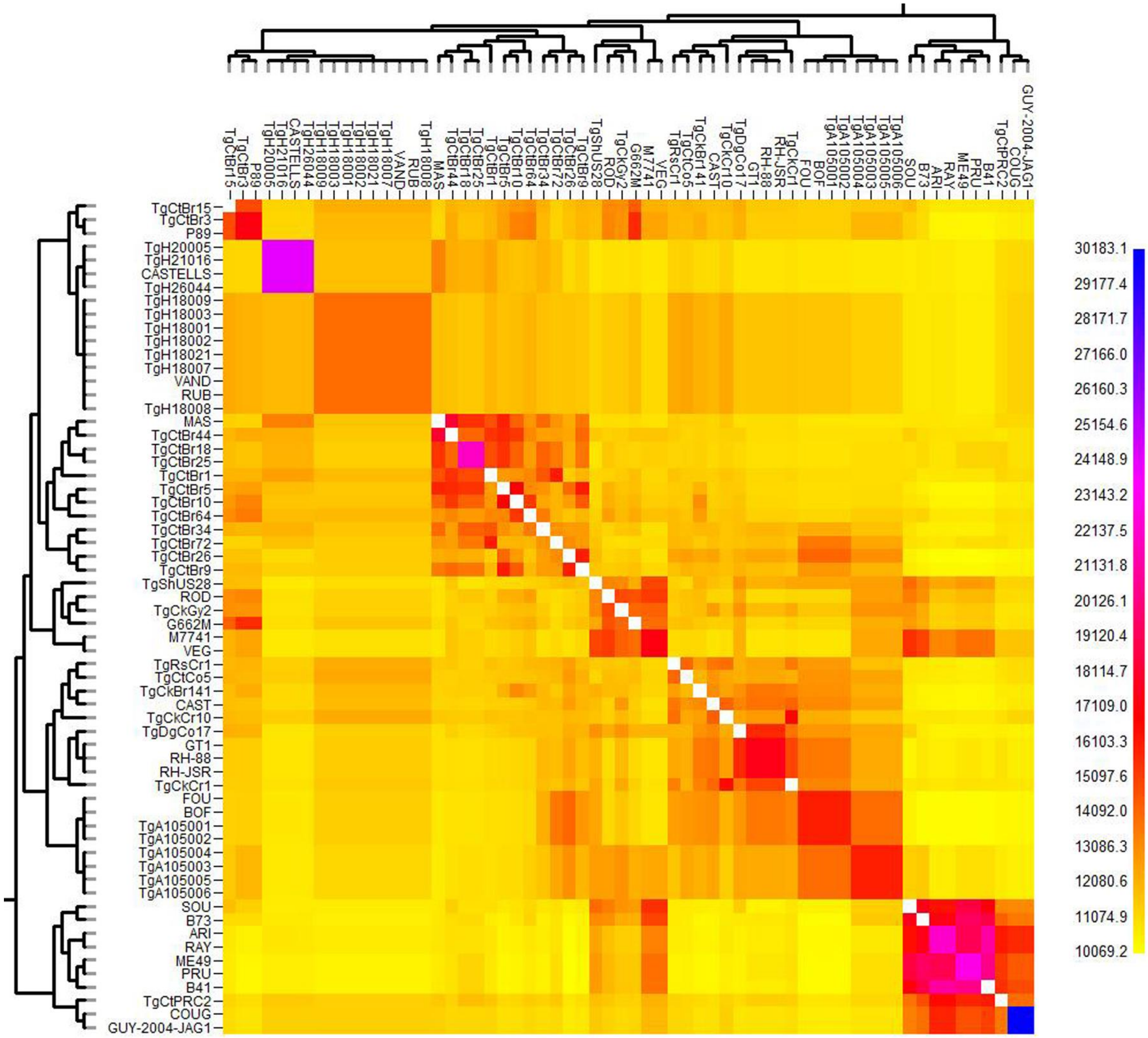


Supplemental Figure 12 Average SNPs per window comparing members of each clade. The average number of SNPs per 10 kb window for all strains within a Clade is plotted: y-axis 0 – 60 (outliers > 60 not plotted) red lines indicate chromosome boundaries. Large regions with low SNP rates for all members within a Clade are indicated by low regions in the plots. For example, chromosome Ia has low SNP rates when comparing all members of Clade B, but not for Clade F. Chromosome numbers are indicated along the x-axis.

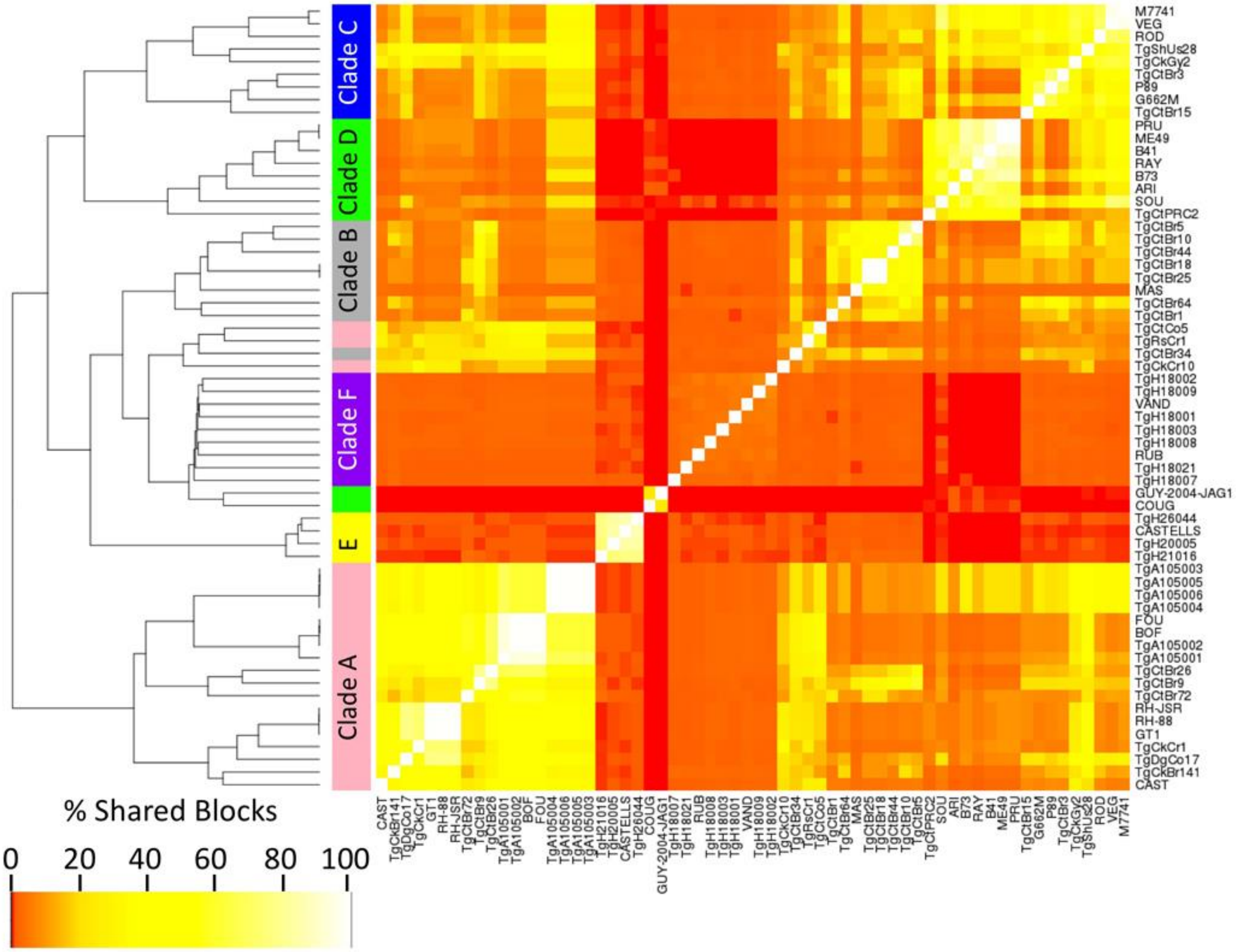


Supplemental Figure 13 Expansion of monomorphic version of ChrIa among *T. gondii* isolates.

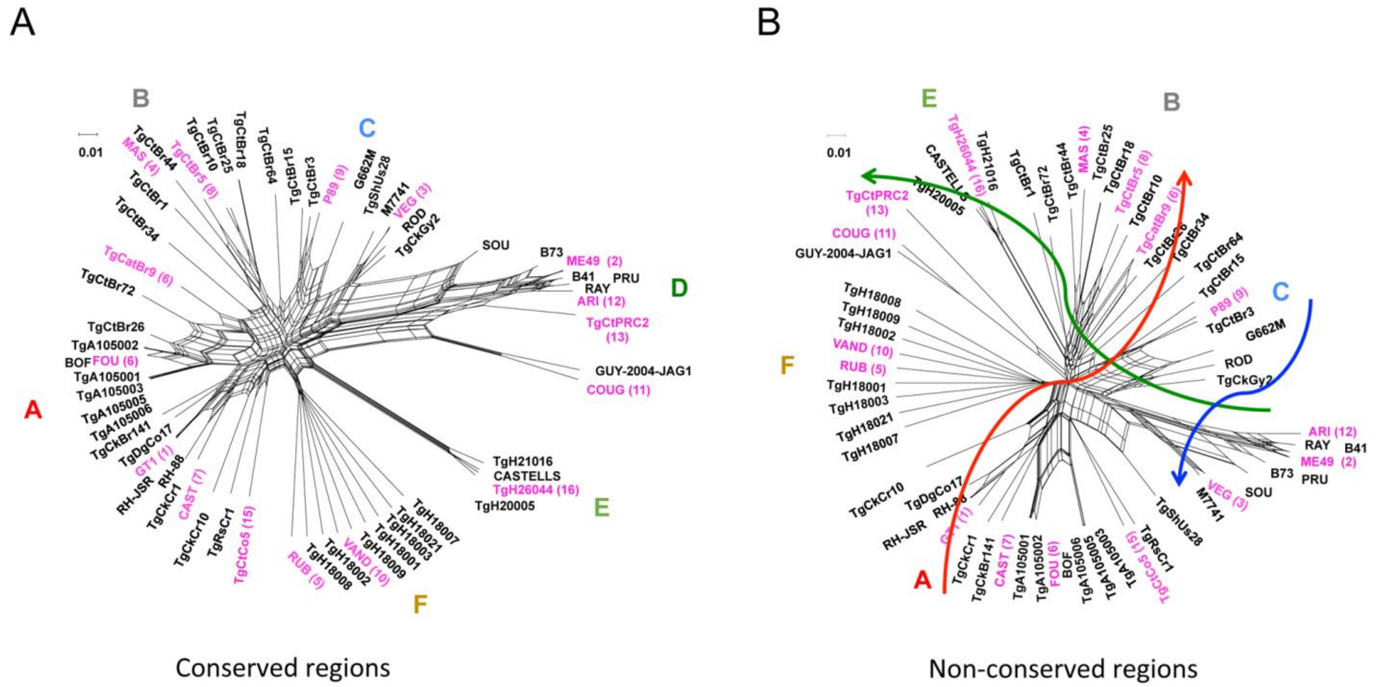
A) Minimum spanning network analysis of ChrIa shows four major clusters; mono-ChrIa, divergent ChrIa, 5’ chimeric-ChrIa, and 3’ chimeric-ChrIa. B) SNP density plots indicate the monophyletic distribution of mono-ChrIa among the majority of *T. gondii* strains. Strains used in each of these plots were defined by minimum spanning network clustering of ChrIa (A). X-axis indicates the relative physical distance of 14 chromosomes of *T. gondii*: y-axis indicated the number of SNPs per 10 kb sliding window.



Supplemental Figure 14 Population structure of *T. gondii* inferred by Bayesian clustering. Pair-wise co-ancestry heatmap reveals the shared haplotypic segments (“chunks”) calculated based on genome wide SNP data of *T. gondii*.



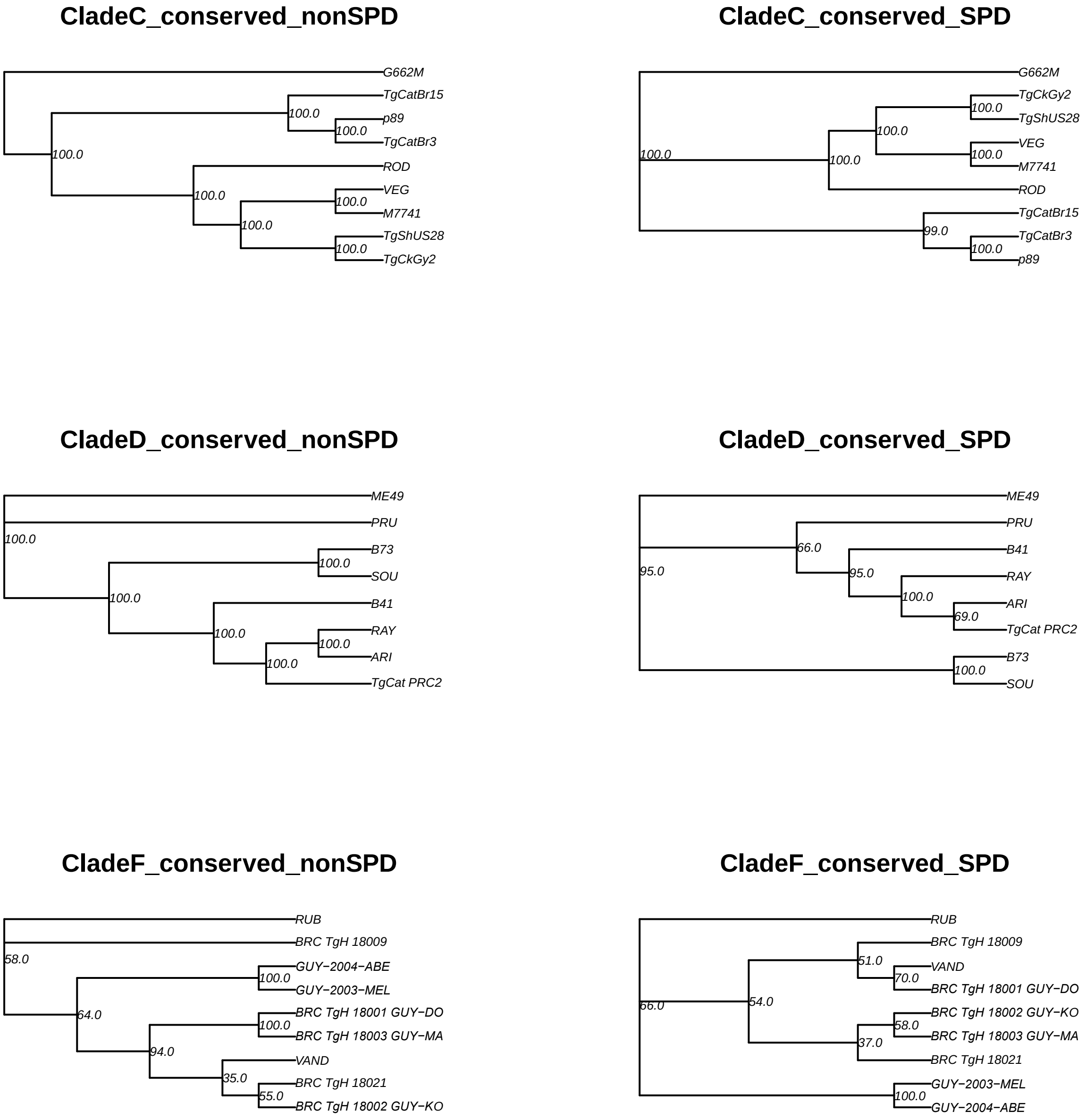
Supplemental Figure 15 Heatmap clustering of co-inheritance of shared blocks. The percentage of shared blocks between two strains was determined for all 62x62 pair-wise strain comparisons (1953 non-redundant comparisons): scale = % shared blocks. Hierarchical clustering on percent shared blocks independently grouped the strains by Clade.



Conserved regions

Non-conserved regions

Supplemental Figure 16 Neighbor-net trees for conserved (A) vs. non-conserved (B) regions of the genome based on total SNPs for the 62 strains. Reference strains for haplogroups (numbers indicated in parentheses) are show in pink. A number of strains group differently in the non-conserved network compared to the conserved network: stains VEG, M7741, and TgShUs28, which are part of clade C, realigned to clade D (blue arrow). Strains TgCtBr9 (6) and TgCtBr72, which are part of clade A, regroup to form part of clade B (red arrow). Finally, strains TgCtPRC2 (1), COUG (11) and GUY-2004-JAG1, which are part of clade D, regroup to a separate long branch (green arrow). Networks were generated as described in Figure 4a. Scale = number of SNPs per site.



Supplemental Figure 17 Phylogenetic trees for SPD and non-SPD genes within conserved regions for Clades C, D, and F. Trees were generated with 100 bootstraps and support values are shown at the branches**.** Trees generated from SPD genes closely resemble those of the non-SPD genes, reflecting their common, recent shared ancestry. Trees were generated using RMxML with the GTR+GAMMA model as further described in the methods.

Supplementary Table 1 Genome assemblies for *T. gondii* reference strains.

	T. gondii GT1	T. gondii VEG	T. gondii ME49	T. gondii MAS	T. gondii RUB	T. gondii CAST	T. gondii TgCtBr5	T. gondii P89	T. gondii VAND	T. gondii COUG	T. gondii ARI	T. gondii TgCtPRC2	T. gondii GAB2-2007-GAL-DOM2	T. gondii TgCtCo5	T. gondii FOU	T. gondii TgCtBr9	H. hammondi H.H.34
Assembly size	64.1 Mbp	63.7 Mbp	65.5 Mbp	61.7 Mbp	62.8 Mbp	63.3 Mbp	62.1 Mbp	62.0 Mbp	62.6 Mbp	64.9 Mbp	63.3 Mbp	63.0 Mbp	63.7 Mbp	63.0 Mb	62.6 Mbp	62.0 Mbp	67.4 Mbp
Number of contings (bp)	3,364	3,424	2,532	6,767	7,495	7,292	9,658	6,415	5,780	15,676	7,609	8,277	7,812	8,277	10,363	7,392	16,398
Contig N50 (bp)	351,608	368,606	1,219,553	38,798	30,445	46,384	26,933	41,694	61,086	33,690	40,798	35,133	59,599	35,133	33,296	35,035	84,429
Number of scaffolds (bp)	313	306	2,266	201	165	127	212	170	2,137	158	126	143	149	143	216	165	14,861
Scaffolds N50 (bp)	4,191,897	3,747,630	6,301,488	1,390,389	2,283,863	3,297,333	1,365,820	1,730,166	1,742,129	3,382,402	2,667,080	1,948,981	2,099,269	1,948,981	1,635,587	1,489,074	1,494,935
Number of fragment 454 reads	NA	NA	1,240,227	NA	1,559,686	NA	NA	NA	2,907,070	NA	NA	NA	NA	2,680,081	NA	NA	3,144,203
Number of 3kbp 454 reads	NA	NA	3,065,409	5,242,673	5,185,312	4,259,026	5,734,612	5,857,993	3,803,275	2,012,690	2,294,377	3,693,418	8,773,773	3,860,620	6,228,112	6,307,240	1,330,431
Number of 8kbp 454 reads	NA	NA	1,182,346	2,362,125	3,111,736	2,856,093	2,233,469	3,038,593	2,465,760	887,134	1,232,621	NA	NA	2,543,092	3,455,458	1,847,984	NA
Number of Illumina PE reads	94,318,960	99,517,042	NA	48,698,504	11,679,625	84,807,782	34,799,463	17,351,409	29,959,353	29,851,932	24,762,190	30,235,296	111,457,054	61,748,554	24,768,278	16,761,892	60,610,959
Number of Sanger reads	1,415,530	1,352,018	980,812	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
Average sequencing depth	67.44x	77.38x	26.55x	42.75x	30.56x	74.24x	31.5x	26.85x	39x	45.9x	46.2x	28.52x	98.92x	53.3x	31.05x	24.5x	66.13x
Accession Number	GCA_000149715.2	GCA_000150015.2	GCA_000006565.2	GCA_000224865.2	GCA_000224805.2	GCA_000256705.1	GCA_000259835.1	GCA_000224885.2	GCA_000224845.2	GCA_000338675.1	GCA_000250965.1	GCA_000256725.1	GCA_000325525.2	GCA_000278365.1	GCA_000224905.2	GCA_000224825.1	GCA_000258005.2

Supplementary Table 2 Synteny Data

Levels of synteny based on orthologous genes (2067 clusters) present in all species

	TG vs HH	TG vs NC	TG vs SN	TG vs ET	HH vs NC	HH vs SN	HH vs ET	NC vs SN	SN vs ET
Total number of syntenic blocks	16	30	95	0	28	95	0	93	0
Total number of genes in syntenic blocks	2115	2064	853	0	2062	862	0	836	0
Number of Mb in syntenic blocks for species 1	59.88	58.67	19.21	0	58.47	19.2	0	18.16	0
Number of Mb in syntenic blocks for species 2	59.49	55.51	35.38	0	55.55	35.82	0	34.72	0
Percent proteome in syntenic blocks for species 1	23.71	23.14	9.56	0	26.14	10.93	0	11.51	0
Percent proteome in syntenic blocks for species 2	26.81	28.41	12.07	0	28.38	12.2	0	11.83	0

Levels of synteny based on orthologous genes (7391 clusters) present in at least two species

	TG vs HH	TG vs NC	TG vs SN	TG vs ET	HH vs NC	HH vs SN	HH vs ET	NC vs SN	SN vs ET
Total number of syntenic blocks	16	46	159	0	44	155	0	145	0
Total number of genes in syntenic blocks	7090	6307	1678	0	6294	1659	0	1525	0
Number of Mb in syntenic blocks for species 1	60.79	62.24	28.42	0	61.83	28.13	0	25.19	0
Number of Mb in syntenic blocks for species 2	60.34	58.75	53.72	0	58.68	54.04	0	50.33	0
Percent proteome in syntenic blocks for species 1	79.48	70.71	18.81	0	79.78	21.03	0	20.99	0
Percent proteome in syntenic blocks for species 2	89.87	86.8	23.75	0	86.62	23.48	0	21.58	0

Materials and Methods

Propagation of strains and isolation of gDNA

Sixty two representative strains of *T. gondii* were selected from different haplogroups from around the world (Dataset 1) [1]. Strains were cultured in human foreskin fibroblast (HFF) cells maintained in DMEM (Invitrogen) containing 10% FBS, 2 mM glutamine, 20 mM HEPES pH 7.5 and 10 µg/ml gentamicin, and harvested after host cell lysis by passing through 3.0 micron polycarbonate filters (Fisher Scientific, UK). Harvested parasites were resuspended in phosphate buffered saline (PBS) at a concentration of approximately 10^7 cell/ml and genomic DNAs were prepared using DNeasy Blood and Tissue kit (Qiagen, USA) according to the manufacturer's instructions.

Genome sequencing and assembly of *T. gondii* reference strains

Improvement of the existing *T. gondii* ME49 assembly was carried out using a combination of Sanger and 454 sequencing technologies. Briefly, one fragment and three paired-end (PE) 454 libraries with 3 kbp (done in duplicate) and 8 kbp inserts were prepared from total genomic DNA (gDNA) extracted from tachyzoite cultures and sequenced in four full plates of a 454 Titanium FLX sequencer. The resulting 5.4 million reads were combined with 980,812 Sanger PE reads from 3 kbp, 10 kbp, and 15 kbp insert libraries generated for the existing genome sequence, screened for contamination and reassembled using Celera Assembler software [2]. Scaffolds were then aligned with MUMmer [3] to *T. gondii* ME49 chromosome sequences from ToxoDB v8.0 to generate super-scaffolds spanning entire chromosomes. The final *T. gondii* ME49 assembly (GenBank ABPA00000000.2, ToxoDB v9.0) is 65.9 Mbp long with an average read depth of 26x and a contig N50 of 1.2 Mbp (Supplementary Table 1).

To reassemble the genome sequences from *T. gondii* GT1 and VEG strains, Illumina PE sequencing libraries with an average insert size of 300 bp were prepared from total gDNA extracted from tachyzoites cultures from each strain and sequenced in one seventh of an Illumina HiSeq 2000 lane. Thereafter, Illumina PE reads (42.5 M for GT1 and 89.7 M for VEG) were examined for contamination, pooled with Sanger PE reads from 3 kbp and 10 kbp libraries (707,774 for GT1 and 676,028 for VEG) from their respective genome assemblies and assembled with Newbler v2.6 [4]. Resulting GT1 and VEG assemblies were deposited in GenBank with accession numbers AAQM00000000.3 and AAYL00000000.2, respectively. Genome sequencing of the remaining reference *T. gondii* strains (Supplementary Table 1) was performed using a 454-Illumina hybrid approach. For each strain up to three 454 gDNA sequencing libraries (one fragment, one 3 kbp PE and one 8 kbp PE libraries) and one Illumina PE library (300 bp inserts) were prepared from gDNA extracted from tachyzoite cultures. Each library was sequenced using either one (fragment or 8 kbp libraries) or two (3 kbp library) full plates of a 454 Titanium FLX run or one half of an Illumina GAII lane. After removing contaminating sequences Illumina and 454 reads were assembled with Newbler v2.6 and the final assembly was submitted to GenBank. Supplementary Table 1 provides assembly statistics as well as GenBank accession numbers.

Genome sequencing of non-reference *T. gondii* strains for SNP discovery

For each non-reference strain, 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 (SRA) at NCBI (Supplementary Dataset 1).

Sequencing of tachyzoite mRNA samples

To aid in the curation of ME49 gene models, two tachyzoite-specific Illumina cDNA 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. Reads were deposited in SRA with accession numbers SRR350746 (ME49) and SRR516419 (GT1).

Structural and functional annotation of the *T. gondii* ME49 genome

Protein sequences from GenBank NR and Pfam-seed databases [5] as well as from the ME49 proteome were aligned to assembled sequences with the utility NAP from the AAT-package [6] to generate protein-based evidence to support identification of gene structures. To train ab-initio gene finders, a tachyzoite-specific RNAseq dataset [7] was obtained from ToxoDB, assembled with Trinity [8] and aligned to the new ME49 assembly sequence with PASA [9]. In addition, a second dataset composed of 136,229 tachyzoite-specific *T. gondii* cDNA/EST Sanger sequences were downloaded from GenBank and then assembled and aligned to the new genome assembly with PASA. This collection of aligned transcripts was used to generate a training set of ~400 manually curated genes, supported by full-length transcripts. GlimmerHMM [10], GeneZilla [10] and Augustus v2.5 [11] were trained and run on the ME49 scaffolds to predict potential protein-coding gene structures. Augustus was run a second time using aligned transcripts as hints to increase prediction accuracy.

The 7,987 gene annotations from the previous assembly were mapped as predictions onto the new genome sequence at three different stringencies with three different sets of programs: (i) MUMmer (nucleotide-based mapping, high stringency), (ii) PASA (nucleotide-based mapping, medium stringency), and (iii) GenWise [12] followed by Geneid [13] (protein-to-nucleotide-based alignment, low stringency). Only three genes, encoding for hypothetical proteins, did not align with any of these methods and hence were discarded. Once all evidence was mapped to the new genome sequence a first set of preliminary working genes models was generated with EVidenceModeler [14] (EVM), a program that predicts gene structures as a weighted consensus of all the evidence available at each particular locus, including transcript and protein alignments and gene predictions. Further improvement of gene structures was performed with PASA by the incorporation of additional transcriptomic evidence from an oocyst-specifc (from the type 2 M4 strain, http://tinyurl.com/mdkdbup) [15] and a bradyzoite-specific (from the type 2 ME49 strain, http://tinyurl.com/ons4zdo) Illumina RNAseq datasets downloaded from ToxoDB. Briefly, each RNAseq dataset was assembled with Trinity and transcripts aligned to the ME49 assembly sequence with PASA. Thereafter, a second functionality of PASA was used to compare and detect annotation inconsistencies between working genes models and mapped transcripts, such as genes that should be created, merged, split or incorporate new exons. A final round of manual curation was performed by the incorporation of structural changes deposited by the scientific community in ToxoDB.

Functional annotation of ME49 protein-coding genes

Predicted peptides from ME49 working models were run through JCVI's autonaming pipeline, that assign product names based on a number of sequence similarity searches whose results are ranked according to a priority rules hierarchy. These analyses included Blastp searches against the previous *T. gondii* ME49 proteome and the GenBank non-redundant protein database, HMM searches against Pfam and TIGRfam [16] databases, and RPS-Blast searches against the NCBI-CDD database [17]. 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.

To predict the possible subcellular localization of predicted peptides, potential signal peptides and transmembrane domains were predicted with signalP and TMHMM programs respectively [18]. Enzyme Commission (EC) numbers were assigned with PRIAM [19] and curated based on annotations deposited by the scientific community in ToxoDB and those kindly provided by Dr. John Parkinson. Gene ontologies were annotated based on pfam2go and ec2go mappings, annotations from the ApiLoc database (http://apiloc.biochem.unimelb.edu.au) and those curated by the research community in ToxoDB.

Assignment of gene pub_locus identifiers

ME49 protein-coding genes from the previous genome assembly inherited a modified version of their pub_locus identifiers by the addition of 100,000 to their number suffixes (for example, pub_locus TGME49_012345 became TGME49_112345 in the new assembly). Annotated tRNAs, rRNAs and newly predicted protein-coding genes were assigned completely new pub_locus identifiers. For the annotation of the remaining *T. gondii* reference genomes and *H. hammondi*, protein-coding genes that could be mapped from the newest ME49 annotation inherited their pub_locus suffix numbers while the rest of the genes acquired new pub_locus identifiers. Partial genes that based on the evidence were split into two or more fragments on different contigs kept the same pub_locus suffix number followed by a letter, different for each fragment.

Domain Identification and characterization 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 [20]. Any protein segment scoring above the trusted cutoff assigned to a particular HMM profile was assigned to that domain. Known domain sequences were then removed from protein sequences and remaining peptides searched against each other using Blastp to identify potential novel domains not represented in Pfam and TIGRfam databases. Similar peptide sequences were clustered by creating a link between any two protein sequences having an identity above 30% over an span of at least 50 residues and an e-value < 0.001. The Jaccard coefficient of community [21] was estimated for each linked pair of peptide sequences *a* and *b* as follows:

$$J_{a,b} = \frac{\#\ \text{distinct accessions matching } a \text{ and } b \text{ including } (a,b)}{\#\ \text{distinct accessions matching either } a \text{ or } b}$$

with the Jaccard coefficient (Ja,b), named link score, providing a measure of similarity between the two proteins. Associations between peptides that had an insufficient link score were eliminated, and the remaining links were used to generate single linkage clusters. Clustered peptides were then aligned using ClustalW to generate conserved protein domains not present in the Pfam and TIGRfam databases. *T. gondii*-specific domain alignments containing three or more proteins were considered true domains for the purpose of building families. Peptide sequences were extracted from alignments and searched back against the *T. gondii* proteome to look for additional members that may have been excluded during earlier stages due to the parameters used. Full-length protein sequences were then grouped on the basis of the presence of Pfam/TIGRfam domains and potential novel domains, named "para" domains. Proteins with exactly the same combination of domains were classified into putative protein families. Gene families having two or more members organized as tandem arrays were identified using an in-house *perl* script.

The top five protein families containing novel para domains (TgFAMs A to E) were selected for further characterization. Briefly, for each family protein sequences were run through Phobius [22] a program that simultaneously identify the presence of potential signal peptides and transmembrane domains. De novo identification of conserved protein domains across members of the same gene family was carried out with MEME [23]. Expression levels for the TgFAMs were obtained from *T. gondii* Affymetrix Array data available from NCBI GEO records GSE32427 [15] and GSE51780. Data were normalized as described in [24].

Annotation of rRNA and tRNA genes

Sequences encoding for 5.8s, 18s and large rRNA subunits were extracted from GenBank entries L25635.1 and L37415.1, aligned to the new assembly with Nucmer [3] and automatically annotated with an in-house *perl* scripts. Transfer RNA encoding genes were predicted with tRNAscan-SE [25].

Annotation of other *T. gondii* reference strains

T. gondii ME49 genes were mapped at high and medium stringency to each *T. gondii* reference assembly with MUMmer and PASA respectively. Genes that mapped without errors with either method were promoted to working gene models while those that failed entered a second round of mapping with EVM. Briefly, coding sequences from failed genes were mapped to the target assembly with gmap [26]. Protein sequences from GenBank NR and Pfam seed databases and from the reannotated ME49 proteome were aligned to contigs with NAP and all RNAseq transcripts assembled with Trinity were mapped with PASA. Ab-initio gene predictions were performed with GlimmerHMM, Genezilla and two runs of Augustus, one of them using gmap-aligned genes as hints. Gene predictions, protein alignments and transcriptomic evidence were then integrated by EVM to annotate remaining working genes on the genome sequences. Lastly, annotated gene structures were updated with PASA based on aligned transcripts and manually inspected.

Functional annotation of protein-coding genes was performed following the same approach as with ME49 with the only exception that genes syntenic to ME49 annotations inherited their product names, Gene ontology (GO) terms and EC numbers while non-syntenic genes acquired their names and other functional annotations from the output of JCVI's autonaming pipeline, PRIAM, ec2go and pfam2go mappings.

Sequencing and assembly of the *Hammondia hammondi* genome

Two 454 sequencing libraries (one fragment and one 3 kbp PE) and one Illumina 300 bp PE library were prepared from total gDNA extracted from *H. hammondi* oocysts (H.H.34 strain). The Illumina library was sequenced from both ends and 100 cycles in one eighth of an Illumina HiSeq 2000 lane while the fragment and 3 kbp PE 454 libraries were sequenced using two and one full 454 plates respectively. Sequencing reads from all libraries were screened for contamination and assembled with Newbler v2.6. The final *H. hammondi* assembly is 67.7 Mbp long with a contig N50 of 84,429 bp and an average sequencing depth of 66x (Supplementary Table 1).

Annotation of the *H. hammondi* assembly

Structural and functional annotations were carried out following a similar approach as the one described for other *T. gondii* reference genomes. In this case, however, GeneZilla, GlimmerHMM and Augustus were trained with a training set composed of 546 *H. hammondi* genes that were manually annotated and whose structures were both conserved in *T. gondii* and full-length supported by *T. gondii* transcripts. In addition, Augustus was run without hints or using evidence from *T. gondii* RNAseq/EST assembled transcripts or ME49 coding sequences.

Annotation of the apicoplast genomes from *T. gondii* reference strains and *H. hammondi*

Structural and functional annotation of the apicoplast rRNA and protein-coding genes were done manually using as reference the annotation of the apicoplast genome from the *T. gondii* RH strain (NC_001799.1). Prediction of apicoplast tRNA genes was performed with tRNAscan-SD.

Estimation of d_N/d_S ratios

Coding sequences from each cluster of orthologous genes from the 16 *T. gondii* reference strains were preprocessed using an in-house *perl* script in order to eliminate redundant sequences and potentially truncated or paralogous genes. Clusters with only one member left after the filtering step were discarded from the analysis. Afterwards, d_N/d_S ratios were calculated using a modified version of the Bioperl script *bp_pairwise_kaks.pl* (http://search.cpan.org/dist/BioPerl/scripts/utilities/bp_pairwise_kaks.pl). Briefly, this script aligns coding sequences in protein space with ClustalW [27] to ensure that codons are aligned properly, projects the alignments back into coding sequences and then estimates the non-synonymous versus synonymous substitution rates based on the maximum likelihood method of Yang [28] implemented in the Codeml software from the PAML package [29]. For this analysis Codeml was run with the following parameters to estimate global d_N/d_S ratios for the entire multiple alignment: runmode = 0, seqtype = 1 and model = 0.

Analysis of orthologous genes using OrthoMCL

Annotated proteomes for *E. tenella* (strain Houghton, release date 2013-11-05), *S. neurona* (strain SN1, release date 2014-09-24), *N. caninum* (strain Liverpool, release date 2011-08-12), *T. gondii* (strain ME49, release date 2013-04-23), and *H. hammondi* (H.H.34 strain, release date 2014-09-03) were analyzed using OrthoMCL v2.0 [30,31] to define orthologous groups using the following parameters: Markov Clustering algorithm mcl ver 12.068 and default parameters (I=1.4, NCBI BLASTP ver 2.2.25 params = -a 1 -F 'm S' -v 100000 -b 100000 -Y 1300000 -e 1e-5).

Determination of Pfam domain abundance in coccidians

The proteomes from *T. gondii* ME49, *H. hammondi*, *N. caninum*, *S. neurona* and *E. tenella* were queried against the Pfam HMM database using HMMER3. Non-overlapping HMM hits above the trusted cutoff spanning at least 50% of the HMM and representing the best hit (lowest e-value) per protein region were selected for further quantification. For every genome, the abundance of each Pfam domain was then estimated as the number of protein sequences having at least one significant hit against a particular Pfam HMM.

Grouping GO annotations of orthologous groups

Clusters of orthologous groups present in the five coccidian genomes *E. tenella*, *S. neurona*, *N. caninum*, *T. gondii*, and *H. hammondi* were identified using OrthoMCL as defined above. These groups were functionally annotated using GO Slim terms, which are designed to group the many different GO terms into smaller groups of related processes (http://geneontology.org/page/go-slim-and-subset-guide). First, the list of GO terms associated with the proteins in the five organisms was inferred from their Pfam assignments, as described above. These terms were mapped onto the smaller-and-broader subset of generic GO Slim terms using map2slim script of the go-perl package, which maps a set of annotations up to their parent GO Slim terms. The GO Slim terms are mapped onto the OrthoMCL groups, and only groups containing the same annotation in >65% of its annotated members are considered.

Mapping non-synonymous SNPs and number of paralogs to iCS382 metabolic model of *T. gondii*

SNPs from the 16 *T*. *gondii* reference strains that correspond to the 382 proteins in the iCS382 metabolic pathway reconstruction of *T*. *gondii ME49* were downloaded from ToxoDB (http://www.toxodb.org/toxo-release4-0/home.jsp) [32]. SNPs were filtered using the following criteria: >80 % of the 16 strains contain an allele (major or minor) that is present in >80% of reads. Further, we considered only non-synonymous coding SNPs (missense and nonsense SNPs) that were present in at least 2 strains to negate the effects of random genetic drift. These criteria resulted in 2,697 nsSNPs. A normalized nsSNP score was generated for each EC number in iCS382 based on the following formula:

$$nsSNPs\ (normalized) = \frac{\sum_{j=1}^{Etot} \frac{number\ of\ nonsynonymous\ SNPs}{protein\ length}}{number\ of\ proteins\ with\ this\ EC\ in\ ME49\ (Etot)}$$

The number of paralogs for an EC was calculated for all 382 proteins from the orthologous groups of 16 reference *T. gondii* strains + *N. caninum* + *H. hammondi* generated using OrthoMCL at an inflation parameter of 1.4.

$$number\ of\ paralogs = \left(\frac{\sum_{j=1}^{Otot} number\ of\ proteins\ from\ 16\ Tg\ strains\ in\ the\ orthomcl\ group}{16} \right) - 1$$

Where *Otot* = number of orthoMCL groups containing a ME49 protein in iCS382 with this EC.

These values are mapped on to the iCS382 metabolic model and visualised using Cytoscape 3.1.1 (http://www.cytoscape.org/) [33].

SNP identification

Illumina reads for each of the 61 other genomes were aligned using Bowtie2 --end-to-end [34] against the ME49 reference genome assembly (release date 2013-04-23). Reads were realigned around gaps using the GATK toolkit [35] and a pileup file was generated using samtools [36]. VarScan [37] pileup2snp was used to make SNP calls with –min-coverage 5, -p-value 0.01 and –min-var-freq 0.8 (80% read consistency). These same parameters were utilized to make like-reference calls for each strain at every position in the ME49 genome where any strain had a SNP call. This identified 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.

Network and principal components analyses

Genome wide single nucleotide polymorphism (SNPs) were saved as FASTA files and directly incorporated into SplitsTree v4.4 [38] to generate unrooted phylogenetic networks using a neighbor-net method and 1,000 bootstrap replicates. Principal components analysis (PCA) was performed by eigenanalysis of a coancestry matrix implemented in fineSTRUCTURE, as described [39].

Chromosome Ia analysis

SNP data for ChrIa were plotted as a minimum spanning tree using SplitsTree v4.4 [38] with 2,000 spring embedded iterations. The 62 strains were clustered into four major groups denoted as monomorphic, divergent, 5'-chimeric, and 3'-chimeric chromosome Ia. SNPs present in each cluster were calculated using a custom script over a 10 kb moving window and plotted using Excel.

Admixture analysis

The population genetic structure of *T. gondii* was determined by an unsupervised clustering algorithm, ADMIXTURE [40] with ancestral clusters set from k = 1 through 10. The number of ancestral clusters k was determined by estimating the low cross-validation error (CV-error) for different k values using 5-fold CV.

Co-ancestry heatmap

We developed a co-ancestry heatmap by using the linkage model of ChromoPainter [41] and fineSTRUCTURE (http://www.paintmychromosomes.com) based on the genome-wide SNP data. For fineSTRUCTURE (version 0.02) [41], both the burn-in and Markov Chain Monte Carlo (MCMC) after the burn-in were run for 10,000 iterations with default settings. Each *T. gondii* strain was considered as a recipient of chunks of DNA donated by the other strains. Inference was performed twice at the same parameter values.

Estimating copy number variation (CNV)

Newly generated genomic sequence reads (454 and Illumina) for 62 strains of *T. gondii*, and previously generated sequences of the ME49 (ABPA00000000.2), GT1 (AAQM00000000.3), and VEG (AAYL00000000.2) strains were downloaded from NCBI (http://www.ncbi.nlm.nih.gov/). The programs sff-dump (NCBI SRA Toolkit) and sffToCA (Celera Assembler) were used to make fastq files from 454 based .sra files, and fastq-dump (NCBI SRA Toolkit) with the spilt-files option was used to make fastq files from Illumina based .sra files. For each strain, the respective .sra files were used to align reads to the 14 ME49 reference chromosomes using Bowtie2 version 2.1.0 [42] with the end-to-end option. The read depth per bp, or read bases (RB), across 8,320 chromosomal-mapped genes was determined using samtools mpileup (SAMtools [36]). The mean of RB ($Gene n$) for base pairs spanning the genomic region of each gene was determined. The baseline 1X RB value for each strain was obtained by calculating the mean of RB ($1Xn$) and standard deviation of RB ($1Xs$) across all base pairs within genes (as one set) in the second and third quartile (4,160 genes) of the $Gene n$ distribution of the 8,320 protein coding genes. The CNV estimate was expressed as

$$CNVestimate = \frac{Gene n}{1Xm}$$

The cutoff for calling copy number variation (CNV) for a gene was based on

$$CNVcutoff = (CNVestimate >= 1Xm + (3*1Xs))$$

Plots were generated in R (http://www.r-project.org/). *T. gondii* gene families organized in tandem arrays were identified with an in-house *perl* script.

Analysis of OrthoMCL species-specific genes

Genes found to be specific to *T. gondii, H. hammondi, or N. caninum* based on OrthoMCL clustering were further analyzed using a combination of sequence alignment tools (**Supplemental Fig. 8A**). First, proteins encoded by species-specific genes were compared against the proteomes of the other two genomes with blastp using an e-value cutoff of $1x10^{-10}$ and at least 50% coverage of the shortest sequence. Genes that passed the cutoff were flagged as “Blastp hit” (see **Supplemental Fig. 8B** and **Supplemental Dataset 6**). Species-specific genes without significant blastp hits were then mapped in nucleotide space against the other two genomes with blastn (e-value ≤ $1x10^{-10}$ and coverage ≥ 70% of the query sequence) and those genes having significant and syntenic hits were selected for further analysis. A blastn hit was considered syntenic if the protein encoded by one of the genes located immediately upstream or downstream of the blastn hit on the subject genome was similar by blastp to the protein encoded by one of the genes at either side of the query sequence (e-value < $1x10^{-10}$ and coverage ≥ 50% of the shortest sequence). Species-specific genes without a significant blastn hit as described above were flagged as “Unique” genes (**Supplemental Fig. 8B**). Coding sequences (CDSs) from species-specific genes mapped by blastn were then aligned to the other two genomes using GMAP (with parameters -n 1 -A -a 1), a splice-aware nucleotide sequence alignment tool [43]. Proteins from species-specific genes whose CDSs did not aligned with GMAP were then mapped to the other two genomes with tblastn (e-value ≤ 1x10^{-5}). GMAP and tblastn hits were manually inspected to determine alignment coverage and the presence of non-sense mutations or frame-shifts in the subject sequence. Species-specific genes (or their protein sequences for tblastn) that mapped with either method with at least 70% coverage and did not present any in-frame STOP codon or frame-shift on the subject sequence were labeled as "Full Length" (**Supplemental Fig. 8B**). Genes (or their proteins) that mapped with less than 70% coverage or presented an in-frame STOP codon or frame-shift on the subject sequence were flagged as "Alternative Gene Model" and constitute potential pseudogenes or functional genes with an altered gene structure compared to the query sequence (**Supplemental Fig. 8B**). Remaining unmapped genes were added to the gene pool flagged as "Unique".

RNA-seq analysis of unique and alternative gene models in *T. gondii*.

Assembled *T. gondii* RNA-seq transcripts generated during the annotation phase of the project were mapped using blastn to CDSs from *T. gondii* genes that were classified as Unique, Alternative Gene Models, BlastP Hits and Full Length based on their comparison to *H. hammondia* and *N. caninum*. Those CDSs that aligned to a transcript across their entire length with at least 95% identity were flagged as supported by RNAseq data. The statistical significance of the difference in relative abundance of each gene category with or without RNAseq support was assessed using the Fisher's exact test with the R function *fisher.test*.

Estimation of maximum alternative allele frequency scores and minimum sequencing depth.

Genome sequencing reads used to assemble the genomes of *H. hammondi* (32,612,714 Illumina reads) and *T. gondii* ME49 strain (4,697,063 454 and Sanger reads) were mapped to their respective genomes with Bowtie2 followed by SNP and INDEL identification with the utilities *mpileup* from *samtools* and *call* from *bcftools* with parameters "-cv -p 1" to ensure that all alternative alleles were reported, independently of their allele frequency. Thereafter, the maximum alternative allele frequency (MAAF) score, defined as the maximum allele frequency reached among the collection of alternative alleles identified in every genomic region of interest, was calculated with an in-house *perl* script. To estimate the minimum sequencing depth (MSD) reached by each locus hit by alternative gene models we calculated the sequencing depth at every nucleotide position of that region with the utility *depth* of the program *samtools* using default parameters and then MSD was calculated using an in-house *perl* script. Binning of MAAF scores and MSD data was carried out with the R program *hist*.

Regions of co-inheritance

To determine the extent of recombination and co-inheritance of blocks between strains, we examined all possible pair-wise comparisons between the 62 strains (62 x 62). There are 3,844 pair-wise comparisons of which 1,953 are unique. The number of SNPs per 10 kb window across the 14 chromosomes for each pair-wise strain comparison was determined. Low SNP regions (regions of recent co-inheritance, or shared blocks) were identified for 10 kb windows that have 5 or fewer SNPs across a continuous stretch of 10 windows (100 kb) allowing for the intermittent outlier. The ratio of windows meeting this criterion out of all 10 kb windows (6,202) was used as the percentage of the genome two strains share as recently co-inherited (referred to as % shared blocks). A heatmap was generated using the R function heatmap.2 (gplots library (http://www.r-project.org/)) with hierarchical clustering on the % shared blocks value. To analyze the composition of genes within shared regions, strains were grouped by Clade based on this hierarchical clustering: Clade A - 18 strains, Clade B - 8 strains, Clade C - 9 strains, Clade D - 8 strains, Clade F - 9 strains. Clade E was not included in this analysis as the strains within this Clade are highly similar. The number of SNPs per 10kb window were averaged for all strains within a Clade, and chromosomal regions with low SNP density were identified as above using 10 kb windows that had 3 or fewer SNPs across a continuous stretch of 10 windows (100 kb), allowing for intermittent outliers.

Identification of SPD genes and clustering within the genome

We identified genes that belong to the SPD families (i.e. *MIC*, *GRA*, *ROP*, *SRS* and *TgFAM*) based on the annotation of ME49 accounting for CNV in determining the gene number. We then mapped the position of the SPDs onto the assembled ME49 genome and defined those that fell into conserved or non-conserved regions. To determine if gene type was independent of region type we compared the observed frequency of SPDs and non-SPD genes in conserved vs. non-conserved regions of the genome using a Chi-squared analysis. The null hypothesis was that the distribution would be random, and there would be no difference between observed and expected. A *P* value of ≤ 0.05 was considered significant cause for rejection of the null hypothesis.

Ancestry of conserved and non-conserved regions

The regions of the genome that are "conserved" were defined as the union of low SNP regions for Clades A, B, C , D, and F. We ignored Clade E because it is highly clonal. From these positions, we separated the SNP matrix (all SNPs for 62 strains) into those that were conserved vs. divergent (non-conserved) for the analyses. We reconstructed phylogenetic trees for the conserved and non-conserved regions using maximum likelihood as implemented in RAxML version 7.3.0 with the GTR+GAMMA model [44]. We calculated the standardized Robinson-Foulds (RF) distance [45] between the conserved and non-conserved trees. The standardized RF distance equals the proportion of negative branches in the two trees. In addition, we generated 500 bootstrap trees for the non-conserved region, and calculated the standardized RF distances between the 500 bootstrap trees and the ML tree for the non-conserved region. Those 500 distances represent the variation of the tree estimate due to mutation when the true tree is the non-conserved tree. We used the maximum distance as the threshold to identify the conserved tree that is significantly different from the non-conserved tree. We also estimated phylogenetic tress for the sequence blocks in conserved and non-conserved regions of Clades C, D, and F (where the null hypothesis for random distribution had been rejected). In each of the conserved and non-conserved regions, genes were separated into SPD and non-SPD genes and separate trees were generated for each using 100 bootstrap replicates using RAxML with the GTR+Gamma model described above. The 100 bootstrap trees were combined into a consensus with support values. Trees were considered congruent if they had no conflicting branches with bootstrap support of > 95%.

Phylogeny

To generate a phylogeny that spans across the Apicomplexa, we chose an ortholog shared by all the major taxa that is defined by OrthoMCL OG5_0126701. The gene id for *T. gondii* is TGME49_249810 and it encodes a 2,749 amino acid protein. In different apicomplexans, this gene is annotated as a DEAD box helicase or activating signal cointegration 1 complex subunit 3. The corresponding protein has regions of high conservation allowing broad phylogenetic comparisons, as well as variable domains that provide better resolution within closely related lineages. Phylogenetic trees were constructed using the Neighbor Joining algorithm with 1,000 Bootstrap replicates as implemented in Geneious ver. 7.1.5 (http://www.geneious.com, [46]) and visualized with FigTree ver. 1.4.0 (http://tree.bio.ed.ac.uk/software/figtree/).

Synteny

Briefly, the OrthoMCL ortholog clusters (see above) were reformatted to represent each pair found in the cluster outside of self-matches. Syntenic blocks were generated between all combinations of genomes as described in [47]. A minimum number of three genes in a 25 kb search window up and downstream from each orthologous gene was required to form a syntenic block. Intervening non-syntenic genes were allowed. Custom scripts were used to calculate the total number of syntenic blocks between genomes as well as the number and percentage of genes present in syntentic blocks. The MCSCAN [48] output was formatted for appropriate input to Circos ver. 0.51 [49] for visualization.

Chromosome painting

Local admixture analyses using an enhanced ADMIXTURE algorithm [40] were conducted on blocks of size 1,000 SNPs on each of the 14 chromosomes of *T. gondii.* Local admixture was used to simultaneously optimize the number of ancestral states for a given genomic region and assign each of the 62 strains to clusters representing these ancestral states. Specifically, for each block of SNPs, we performed local admixture analysis with the number of ancestral states k = 2-7, and then chose the one with the minimum cross-validation (CV) error as the optimal number of ancestral states. The ".Q" output file for the optimal ancestral states provided the probability of each strain being assigned to each ancestral state. If the probability was greater than 0.9, the strain was assigned to the corresponding ancestral state. For each sequence block present in an ancestral state cluster, we colored the region according to clades represented in Figure 4a based on majority rule (*i.e*. we counted how many sequences in the ancestral state cluster belonged to 1 or more of the 6 groups in Figure 4a and we assigned the color that represented the largest number of sequences to all of the sequences in that ancestral state cluster).

References

[1] 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).

[2] Myers, E. W. *et al.* A whole-genome assembly of Drosophila. *Science* **287**, 2196-2204(2000).

[3] Delcher, A. L., Phillippy, A., Carlton, J. & Salzberg, S. L. Fast algorithms for large-scale genome alignment and comparison. *Nucleic Acids Res* **30**, 2478-2483(2002).

[4] Miller, J. R., Koren, S. & Sutton, G. Assembly algorithms for next-generation sequencing data. *Genomics* **95**, 315-327(2010).

[5] Sonnhammer, E. L., Eddy, S. R. & Durbin, R. Pfam: a comprehensive database of protein domain families based on seed alignments. *Proteins* **28**, 405-420(1997).

[6] Huang, X., Adams, M. D., Zhou, H. & Kerlavage, A. R. A tool for analyzing and annotating genomic sequences. *Genomics* **46**, 37-45(1997).

[7] Reid, A. J. *et al.* Comparative genomics of the apicomplexan parasites *Toxoplasma gondii* and *Neospora caninum*: Coccidia differing in host range and transmission strategy. *PLoS Pathog* **8**, e1002567(2012).

[8] Grabherr, M. G. *et al.* Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat Biotechnol* **29**, 644-652(2011).

[9] Haas, B. J. *et al.* Improving the Arabidopsis genome annotation using maximal transcript alignment assemblies. *Nucleic Acids Res* **31**, 5654-5666(2003).

[10] Majoros, W. H., Pertea, M. & Salzberg, S. L. TigrScan and GlimmerHMM: two open source ab initio eukaryotic gene-finders. *Bioinformatics* **20**, 2878-2879(2004).

[11] Stanke, M., Tzvetkova, A. & Morgenstern, B. AUGUSTUS at EGASP: using EST, protein and genomic alignments for improved gene prediction in the human genome. *Genome Biol* **7** **Suppl 1**, S11 11-18(2006).

[12] Birney, E., Clamp, M. & Durbin, R. GeneWise and Genomewise. *Genome Res* **14**, 988-995(2004).

[13] Blanco, E., Parra, G. & Guigo, R. Using geneid to identify genes. *Curr Protoc Bioinformatics* **Chapter 4**, Unit 4 3(2007).

[14] Haas, B. J. *et al.* Automated eukaryotic gene structure annotation using EVidenceModeler and the Program to Assemble Spliced Alignments. *Genome Biol* **9**, R7(2008).

[15] Fritz, H. M. *et al.* Transcriptomic analysis of toxoplasma development reveals many novel functions and structures specific to sporozoites and oocysts. *PLoS One* **7**, e29998(2012).

[16] Haft, D. H., Selengut, J. D. & White, O. The TIGRFAMs database of protein families. *Nucleic Acids Res* **31**, 371-373(2003).

[17] Marchler-Bauer, A. *et al.* CDD: conserved domains and protein three-dimensional structure. *Nucleic Acids Res* **41**, D348-352(2013).

[18] Emanuelsson, O., Brunak, S., von Heijne, G. & Nielsen, H. Locating proteins in the cell using TargetP, SignalP and related tools. *Nat Protoc* **2**, 953-971(2007).

[19] Claudel-Renard, C., Chevalet, C., Faraut, T. & Kahn, D. Enzyme-specific profiles for genome annotation: PRIAM. *Nucleic Acids Res* **31**, 6633-6639(2003).

[20] Eddy, S. R. Accelerated Profile HMM Searches. *PLoS Comput Biol* **7**, e1002195(2011).

[21] Jaccard, P. The Distribution of the Flora in the Alpine Zone. *The New Phytologist*, 37-50(1912).

[22] Kall, L., Krogh, A. & Sonnhammer, E. L. A combined transmembrane topology and signal peptide prediction method. *J Mol Biol* **338**, 1027-1036(2004).

[23] Bailey, T. L. *et al.* MEME SUITE: tools for motif discovery and searching. *Nucleic Acids Res* **37**, W202-208(2009).

[24] 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).

[25] Lowe, T. M. & Eddy, S. R. tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Res* **25**, 955-964(1997).

[26] Wu, T. D. & Watanabe, C. K. GMAP: a genomic mapping and alignment program for mRNA and EST sequences. *Bioinformatics* **21**, 1859-1875(2005).

[27] Thompson, J. D., Gibson, T. J. & Higgins, D. G. Multiple sequence alignment using ClustalW and ClustalX. *Curr Protoc Bioinformatics* **Chapter 2**, Unit 2 3(2002).

[28] Goldman, N. & Yang, Z. A codon-based model of nucleotide substitution for protein-coding DNA sequences. *Mol Biol Evol* **11**, 725-736(1994).

[29] Yang, Z. PAML: a program package for phylogenetic analysis by maximum likelihood. *Comput Appl Biosci* **13**, 555-556(1997).

[30] Fischer, S. *et al.* Using OrthoMCL to assign proteins to OrthoMCL-DB groups or to cluster proteomes into new ortholog groups. *Curr Protoc Bioinformatics* **Chapter 6**, Unit 6 12 11-19(2011).

[31] Li, L., Stoeckert, C. J., Jr. & Roos, D. S. OrthoMCL: identification of ortholog groups for eukaryotic genomes. *Genome Res* **13**, 2178-2189(2003).

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

[33] Shannon, P. *et al.* Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res* **13**, 2498-2504(2003).

[34] Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat Methods* **9**, 357-359(2012).

[35] DePristo, M. A. *et al.* A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat Genet* **43**, 491-498(2011).

[36] Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078-2079(2009).

[37] Koboldt, D. C. *et al.* VarScan: variant detection in massively parallel sequencing of individual and pooled samples. *Bioinformatics* **25**, 2283-2285(2009).

[38] Huson, D. H. & Bryant, D. Application of phylogenetic networks in evolutionary studies. *Mol. Biol. Evol.* **23**, 254-267(2006).

[39] Price, A. L. *et al.* Principal components analysis corrects for stratification in genome-wide association studies. *Nat Genet* **38**, 904-909(2006).

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

[41] Lawson, D. J., Hellenthal, G., Myers, S. & Falush, D. Inference of population structure using dense haplotype data. *PLoS Genet* **8**, e1002453(2012).

[42] Langmead, B., Trapnell, C., Pop, M. & Salzberg, S. L. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol* **10**, R25(2009).

[43] Wu, T. D. & Wantanabe, C. K. GMAP: a genomic mapping and alingment program for mRNA and EST sequences. *Bioinformatics*, 1859-1875(2005).

[44] Stamatakis, A. RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics* **22**, 2688-2690(2006).

[45] Robinson, D. R. & Foulds, L. R. Comparison of phylogenetic trees. *Mathmatical Biosciences* **53**, 131-147(1981).

[46] Kearse, M. *et al.* Geneious Basic: an integrated and extendable desktop software platform for the organization and analysis of sequence data. *Bioinformatics* **28**, 1647-1649(2012).

[47] DeBarry, J. D. & Kissinger, J. C. Jumbled genomes: missing Apicomplexan synteny. *Mol Biol Evol* **28**, 2855-2871(2011).

[48] Tang, H. *et al.* Unraveling ancient hexaploidy through multiply-aligned angiosperm gene maps. *Genome Res* **18**, 1944-1954(2008).

[49] Krzywinski, M. *et al.* Circos: an information aesthetic for comparative genomics. *Genome Res* **19**, 1639-1645(2009).