Published in final edited form as:

Curr Opin Microbiol. 2018 December ; 46: 86–92. doi:10.1016/j.mib.2018.09.001.

Effector variation at tandem gene arrays in tissue-dwelling coccidia: who needs antigenic variation anyway?

Matthew L. Blank¹ and Jon P. Boyle^{1,*}

¹Department of Biological Sciences, University of Pittsburgh, Pittsburgh, Pennsylvania, USA

Abstract

Locus expansion and diversification is pervasive in apicomplexan genomes and is predominantly found in loci encoding secreted proteins that interact with factors outside of the parasite. Key for understanding the impact of each of these loci on the host requires identification and functional characterization of their protein products, but these repetitive loci often are refractory to genome assembly. In this review we focus on *Toxoplasma gondii* and its nearest relatives to highlight the known impact of duplicated and diversified loci on our understanding of the host-pathogen molecular arms race. We describe current tools used for the identification and characterization of these loci, and review the most recent examples of how gene-expansion driven diversification can lead to novel gene functions.

Keywords

standing evolution; gene duplication; *Toxoplasma gondii*; gene diversification; evolution; MAF1; *Neospora caninum*

Background

Locus duplication events are an essential driver of genetic innovation, particularly with respect to the emergence of novel traits [1–5]. In the event that a given gene is perfectly duplicated, the most common fate is maintenance at low allele frequency or loss on the population level due to neutral impact, depending on the effective population size [6, 7]. However, in cases where gene dosage leads to improved fitness, the altered allele can be maintained in the population as either standing genetic variation or become fixed via drift or selective sweep [5, 8, 9]. If increases in gene dosage remain advantageous the additional gene copy can be maintained under the same selective constraints as the original copy [8, 9]. However, selection-driven diversification can result in more dramatic changes in gene function, including sub- and neo-functionalization [3, 4], where functions are distributed

*Corresponding author: Jon P. Boyle, PhD, LSA101A, 4249 Fifth Avenue, Pittsburgh, PA 15260, Phone: 412.624.5842, Fax: 412.624.4759, boylej@pitt.edu.

Conflicts of interest

None.

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

across distinct copies or new functions arise. While the importance of Gene Expansion-Driven Diversification (GEDD) was known long before complete genome sequences were made available for any organism, in the post-genomic era structural variation analysis forms a consistent part of any genome-wide analysis, and variations in copy number have been extensively linked to a variety of phenotypes in humans and other organisms [3, 9–12].

GEDD is an appealing mechanism for virulence factor evolution in pathogens, given that duplicate gene copies can be used as fodder in the molecular arms race to counter host evolutionary adaptations. In Apicomplexans like *Plasmodium falciparum* GEDD has clearly played a role in the emergence of the expanded *var* family of genes which are the primary loci mediating antigenic variation in this organism [13, 14]. Outside of the Apicomplexa, the *T. brucei* genome contains a vast pool of extensively duplicated variant surface glycoprotein (VSG) genes [15] that have emerged via GEDD and expression of single of a small number of VSG genes underlie the canonical waves of parasitemia associated with *T. brucei* infection [16]. These classic examples demonstrate how GEDD can provide the tools required to avoid parasite clearance by host adaptive immune responses, promoting parasite survival and ultimately parasite transmission.

We are very interested in the role that GEDD has played in the emergence of either refined or novel gene functions in apicomplexan parasites like *Toxoplasma gondii*, *Hammondia hammondi* and *Neospora caninum*, tissue-dwelling coccidians responsible for billions of infections worldwide [17, 18]. In contrast to *T. brucei* and *Plasmodium spp.*, tissue-dwelling coccidians like *T. gondii* do not appear to have a sophisticated means of antigenic variation. However, our work and that of others [10, 12, 19–21] has shown that these parasite species do have a large collection of effectors, many of which have undergone extensive GEDD. Our work has also shown that the impact of these loci on infection outcome is most often NOT due to changes in gene dosage or temporal variation in expression, but rather on selection-driven changes in protein function across individual gene paralogs that are all expressed simultaneously. Others have thoroughly examined the extent of gene family expansion in *T. gondii* and other organisms and we refer the reader to these excellent papers if they are interested in gene duplicates that are distributed throughout the genome (e.g., [19, 22]). In this review we discuss recent examples of the impact of GEDD on *T. gondii* interactions with its host, with an exclusive focus on *local* tandem gene expansions followed by diversification that have led to refined and/or completely distinct functions in individual paralogs.

Genome-wide characterization of expanded loci in *T. gondii* and its relatives

T. gondii genomics is very advanced. The genome for this parasite as well as many of its near and far relatives have been sequenced and extensively annotated, including the nearest extant relative of *T. gondii*, *Hammondia hammondi* [19, 23], and the less-related cattle parasite *Neospora caninum* [22]. These annotated assemblies are available for download and query on ToxoDB.org [24]. We and others [10, 19] have used raw genome sequence data from these sequencing projects (mostly Sanger technology for *T. gondii*/*N. caninum* and a combination of 454 and Illumina short-read data for others) to identify genomic regions likely to contain tandem gene expansions, as evidenced by increased sequence coverage.

After removal of low complexity reads with short repeats, we are then able to identify tandem gene arrays in multiple genomes, as we did for *T. gondii*, *N. caninum*, and *H. hammondi* [10], exploiting the high level of gene-by-gene synteny in these organisms. This pipeline was crucial for identification of loci with increased coverage in *T. gondii* compared to *N. caninum* and *H. hammondi*, and facilitated downstream studies aimed at understanding the functional impact of GEDD on the evolution of novel traits in this organism [10].

Our work and that of others thus far clearly shows that tandemly expanded loci are quite common in *T. gondii* and its near relatives and enriched for genes with the capacity to encode factors that work outside of the parasite rather than within it [10, 19]. During and after host cell invasion, *T. gondii* and its relatives secrete a suite of effector proteins from specialized secretory organelles [25, 26]. Forty-two of the 53 identified expanded loci identified in *T. gondii* were predicted secretory proteins and 29 contained an Nterminal signal sequence [10]. Some of these loci had been previously annotated and studied because of their significant impact on pathogen virulence and host interactions. Members of the *ROP2* superfamily like *ROP5*, *ROP2* and *ROP4/7* are tandemly expanded [25, 27–29]. In most cases the impact of diversification in individual paralogs is unclear, with the exception of *ROP5* (discussed below), where there is clear evidence linking locus-specific diversification to gene function [29, 30]. *T. gondii* dense granule proteins like Mitochondrial Association Factor 1a,b (MAF1a,b) mediate the bulk of their effects on the host-cytoplasmic side of the parasitophorous vacuole [12, 31]. The *SAG*-related surface antigen family (*SRS* genes) are expressed at the parasite surface and have significant levels of local, tandem GEDD in both *T. gondii* and *N. caninum*. Select members of this family are known to play a critical role in host immune evasion [10, 32], but little else is known about their contribution to host interactions and/or parasite biology. Even though the majority of these expanded loci do not have a predicted function, they provide a curated foundation to identify potential and exciting host-pathogen genes for study [10].

In contrast to extensive tandem locus expansion at host-interacting loci, examples of duplication of genes encoding cytoskeletal constituents or metabolic enzymes are more rare, likely due to tighter constraints on gene dose [19]. One exception to this can be found in two multicopy gene cassettes encoding inner membrane complex (IMC) proteins which localize to the apical complex of *T. gondii* and maintain rigidity of the conoid structure [26, 33, 34]. Two of these proteins (IMC4 and IMC1; TGME49_231630 and TGME49_231640, respectively) are derived from a gene duplication/inversion event on Chromosome VIII such that, unlike most tandem gene arrays they are found in a head-to-head orientation. Interestingly, the IMC8 and IMC5 genes are arranged in a perfectly analogous fashion on chromosome X (TGME49_224520 and TGME49_224530)

suggesting that there was an ancestral duplication and translocation event at this locus that predated *N. caninum*, *H. hammondi* and *T. gondii* [19]. Absent their role in cytoskeletal rigidity the significance of the diversification across these IMC gene paralogs is unclear, although given their deep ancestry it is highly likely that they have evolved refined or even novel functions within the IMC. The NTPase locus is also characterized by a tandem duplication (TGME49_277240 and TGME49_277270) and the proteins encoded by these genes (NTPases I and II) are 59% identical to one another [19]. This duplication is difficult

to trace across other *T. gondii* strains and related species, likely due to issues with poor assembly. Interestingly, these NTPases are not cytoplasmic in the parasite but rather are secreted into the parasite-containing vacuole. Their role in *T. gondii* biology is unclear, in that they do not affect virulence in mice and do not subvert the murine immune response [35].

Expanded loci in *T. gondii* and its close relatives are often found at the host-parasite interface

Secreted proteins from the rhoptry organelles of *T. gondii*, *N. caninum* and *H. hammondi* have been extensively studied because this class of proteins play key roles in pathogen virulence, particularly ROP18 and ROP5 (both members of the ROP2 superfamily) [25, 27, 36]. ROP5 is a pseudokinase that traffics to the PVM [28, 29] and is critical for *T. gondii* virulence. *T. gondii* harbors a diverse and expanded *ROP5* locus that has diverged both within and between strains [29], making it one of the more dramatic examples of tandem locus expansion followed by diversification. Within each strain, the *ROP5* locus encodes three distinct isoform clades based on amino acid sequences: A, B, and C [28, 29], suggesting that duplication was followed by selection-driven diversification at this locus [29]. In support of this, ectopic expression of individual ROP5 paralogs in *ROP5* null parasites and infection into mice show full recovery of the virulence phenotype only after complementation with both a ROP5A and B paralogs (but conversely not with two copies of paralog A) [29]. Based on quantitative genetic data the type I and III alleles are much more potent with respect to virulence in the mouse compared to the type II allele [29, 30, 37], although the root cause of this at the sequence level for this is unclear. Given the diversity present at this locus it is likely that multiple factors determine the phenotypic differences between these loci, including the exact composition of the locus with respect to the A, B and C paralogs.

It is not just *T. gondii* that has a set of species-specific GEDD loci that are involved in mediating interactions with the host. Our analysis on the *N. caninum* genome identified uniquely expanded loci in this species as well [10], and we observed a remarkable lack of overlap in the putative biological function of expanded loci in this species compared to *T. gondii*. Specifically, nearly half of the expanded loci predicted to encode secreted proteins were clear members of the SRS superfamily of surface antigens [10, 22]. SRS proteins play a major role in parasite attachment and invasion [38], they are developmentally expressed, often the target of host antibody responses, and in some cases are required for establishing latent infections in the host [32, 39]. While SRS subfamilies are scattered throughout the genome (similar to members of the ROP2 superfamily), in *N. caninum* they are extensively duplicated in tandem [10, 21]. For example, the *NCLIV_002220* locus has five tandemly duplicated and diversified copies (ranging from 57–88% similar identities) as compared to *T. gondii* maintaining only one copy [10, 22]. While *N. caninum* harbors clusters of expanded SRS genes throughout the genome, it is rare that each copy is expressed in comparison to *T. gondii* [22]. The significance of certain copies being on or off remains elusive, but a proposed hypothesis describes that this differential expression leads to the restricted host range of *N. caninum* [22].

Neofunctionalization in the locus driving interactions between the parasite and host mitochondria

The expanded and diversified gene, *mitochondrial association factor 1 (MAF1)*, provides an exciting example of the characterization of an expanded locus identified originally by QTL mapping and screening for expanded loci [10]. The *MAF1* locus (*TGME49_220950* and *TGME49_279100*) drives expression of unique classes of MAF1 paralogs ranging in 59% to 95% similar identity with one another.

MAF1 copy number varies across strain and species [10]. It is estimated that *N. caninum* harbors 1 or 2 identical copies of *MAF1* while *H. hammondi* has 2 diverse copies suggesting that a duplication and diversification event occurred during *Hammondia* speciation (Fig 1) [12]. *T. gondii* bears 6–10 copies of MAF1 depending on strain type (Fig 1) [12]. Some *MAF1* copies mediate the characteristic phenotype of host mitochondrial association (HMA) [12]. HMA has been recognized for quite some time as the intimate associate of the host mitochondria with the parasite-containing parasitophorous vacuole membrane (PVM) [31, 40–42]. *T. gondii* types I and III are HMA-competent while Type II is HMA-incompetent [12]. The MAF1a and MAF1b paralogs were assembled into clades based on the presence (MAF1b) or absence (MAF1a) of a proline rich region between the putative transmembrane domain and C-terminus of the protein [12]. Interestingly, paralogs belonging to the MAF1b paralog clade are solely capable of complementing HMA in Type II *T. gondii* to near wildtype Type I levels while MAF1a paralogs do not [12]. These data suggest the requirement of only the MAF1b paralogs in driving HMA [12]. Important to this finding is that *N. caninum* is HMA-incompetent and harbors only a MAF1a copy while *H. hammondi* is HMA-competent and has both MAF1a and MAF1b.

Crystallographic structures reveal that the MAF1a and MAF1b paralogs are highly similar to one another in their C-termini and adopt an α/β globular structure homologous to ADP-ribose (ADPr) binding macro-domains [11]. MAF1a and MAF1b bind ADPr with micromolar affinity, suggesting this site remains as a functional structure on both MAF1 paralogs (Fig 1) [11]. While the role of this macrodomain is unknown to date, there is no direct evidence linking HMA to ADPr coordination [11].

In contrast to the macrodomain, there are other regions in the solved MAF1 structure that are divergent. One such region is a charged basic patch in the last 16 Cterminal residues of MAF1a (Fig 1) [11, 12]. Upon mutation of the aligned uncharged residues in MAF1b to the charged residues in MAF1a, HMA is completely lost (Fig 1) [11]. The C-terminus of MAF1 encompassing these key residues has acquired significant diversity between the two paralog clades over time while the ADPr site has maintained somewhat similar binding pockets and is shared with the ancestral MAF1 paralog (Fig 1) [11]. This work suggests a modularity to MAF1 where only the components driving HMA were selected for over time, highlighting the significance of this neofunctionalized expanded locus and its evolutionary role in exploiting the host [11].

GEDD loci in *T. gondii* and its near relatives: current annotation and sequencing status

While *de novo* genome assemblies exist for 63 *T. gondii* strains there are multiple gaps in each of these assemblies. For example, *T. gondii* ME49 genome (v30; www.toxodb.org) has 224 gaps distributed across the 14 chromosomes. On the 5 largest chromosomes almost 50% (36/79) of the gaps are within 100,000 bp of the nearest chromosome end, while the remaining gaps are evenly spaced across the chromosomes (Fig 2a,b). Closer inspection shows that many of the subtelomeric gaps are immediately adjacent to known tandem gene expansions, including *ROP5* and *ROP42* (Fig 2a), consistent with tandem expansions causing a breakdown in sequence assembly. Interestingly, in the ME49 strain the *MAFI* locus is not near any sequence assembly gaps (Fig 2a), however in the VEG strain there are multiple sequence gaps near this locus (e.g., *TGVEG_279100A*; www.toxodb.org). These data clearly show that data are missing at tandemly expanded effector loci known to play a role in *T. gondii* pathogenesis. Given our interest in *T. gondii* secreted effectors that are not only duplicated but have also undergone selection-driven diversification, the current genomic assembly does not permit us to directly compare individual paralog sequences since gene models predicted in the current genome are almost certainly chimeras of raw sequence reads from distinct paralogs. In some cases (as for the *T. gondii* *ROP5* locus; [30, 37]) it is possible to manually reassemble the raw sequence reads that map to the repetitive locus and separate them into putative “real” paralogs. However, this approach is not amenable to high throughput, and does not work for all loci depending on how the complexity of the structural variation and the degree of divergence across paralogs. So at the present moment, current sequencing technology fails to capture the extensive diversity in gene sequence that may be present at any number of tandemly expanded loci. If *T. gondii* *ROP5* and *MAFIa,b* genes are representative, the sequences missing from current chromosome-sized sequence assemblies are likely to be important sources of unappreciated effector variation for *T. gondii* as well as other pathogen species.

Conclusion and perspective

The examples above highlight a large body of work describing GEDD-driven protein products involved in the host-pathogen molecular arms race. The *MAFI* locus is an excellent example of the importance of GEDD, in that it harbors paralogs that have maintained an ancestral macro domain while only certain paralogs are endowed with the ability to mediate host mitochondrial association (HMA) [11, 12]. In this case GEDD led to the emergence of a novel function [12]. Similarly, the *ROP5* locus harbors 3 broad classes of paralogs which are synergistic in their impact on *T. gondii* virulence in mice [20]. Whether this is due to novel functions in each of the paralog families is unknown, but the impact of GEDD in the emergence of these paralog families within the tandem gene cluster is quite clear. The total number of loci that have undergone GEDD is difficult to determine since tandem gene arrays are almost always poorly assembled, but there are likely between 20–30 in the *T. gondii* genome alone [10]. The emergence of single molecule long read technologies such as PacBio SMRT cells and Oxford Nanopore DNA sequencing holds great promise to determine the exact sequences of these tandem gene arrays so that the precise amount of sequence variation between paralogs can be determined. As these methods are applied we anticipate refining our understanding of the extent of GEDD at loci like *MAFI* and *ROP5*,

and also that GEDD is equally pervasive at other loci. This will greatly facilitate our efforts at better understanding the role of such loci in tissue dwelling coccidians like *T. gondii* and its near relatives, specifically with respect to how individual paralogs might differ in their interactions with the host.

Acknowledgements

Research in the authors' laboratory was funded by the National Institutes of Health (NIH) AI114655 to JPB. Our thanks the authors of work cited in this review and apologies to authors whose work was not cited because of length constraints.

References

* of special interest

** of outstanding interest

1. Sudmant PH, Kitzman JO, Antonacci F, Alkan C, Malig M, Tsalenko A, et al. Diversity of human copy number variation and multicopy genes. *Science*. 2010;330(6004):641–6. [PubMed: 21030649]
2. Campbell CD, Sampas N, Tsalenko A, Sudmant PH, Kidd JM, Malig M, et al. Population genetic properties of differentiated human copy-number polymorphisms. *Am J Hum Genet*. 2011;88(3):317–32. [PubMed: 21397061]
3. Ohno S *Evolution by gene duplication* London, New York,: Allen & Unwin; Springer-Verlag; 1970 xv, 160 p. p.
4. Espinosa-Cantu A, Ascencio D, Barona-Gomez F, DeLuna A. Gene duplication and the evolution of moonlighting proteins. *Front Genet*. 2015;6:227. [PubMed: 26217376]
5. Magadum S, Banerjee U, Murugan P, Gangapur D, Ravikesavan R. Gene duplication as a major force in evolution. *J Genet*. 2013;92(1):155–61. [PubMed: 23640422]
6. Koonin EV. Splendor and misery of adaptation, or the importance of neutral null for understanding evolution. *BMC Biol*. 2016;14(1):114. [PubMed: 28010725]
7. Lynch M, Conery JS. The evolutionary fate and consequences of duplicate genes. *Science*. 2000;290(5494):1151–5. [PubMed: 11073452]
8. Selection Wagner A. and gene duplication: a view from the genome. *Genome Biol*. 2002;3(5):reviews1012.
9. Reams AB, Roth JR. Mechanisms of gene duplication and amplification. *Cold Spring Harb Perspect Biol*. 2015;7(2):a016592. [PubMed: 25646380] *[9]: Reams, et al. 2015 This paper extensively describes the molecular mechanisms that give rise to gene duplication and locus amplification events in both microbial and metazoan organisms.
10. Adomako-Ankomah Y, Wier GM, Borges AL, Wand HE, Boyle JP. Differential locus expansion distinguishes *Toxoplasmatinae* species and closely related strains of *Toxoplasma gondii*. *MBio*. 2014;5(1):e01003–13.
11. Blank ML, Parker ML, Ramaswamy R, Powell CJ, English ED, Adomako-Ankomah Y, et al. A *Toxoplasma gondii* locus required for the direct manipulation of host mitochondria has maintained multiple ancestral functions. *Mol Microbiol*. 2018.**[11]: Blank et al., 2018 This study describes the evolution of a *T. gondii*-specific tandemly expanded locus, MAF1, that has undergone GEDD. The solved crystal structure of both the ancestral and neofunctionalized MAF1 paralogs revealed strong selective pressure on the MAF1 Cterminus, critical for the gain of host mitochondrial association while maintaining an ancestral ADPr binding cleft.
12. Adomako-Ankomah Y, English ED, Danielson JJ, Pernas LF, Parker ML, Boulanger MJ, et al. Host Mitochondrial Association Evolved in the Human Parasite *Toxoplasma gondii* via Neofunctionalization of a Gene Duplicate. *Genetics*. 2016;203(1):283–98. [PubMed: 26920761] **[12]: Adomako-Ankomah and English, et al., 2017 Key for understanding the importance of tandem expansion at a given locus is identifying the precise function of each paralog. This study identified the divergent host mitochondrial association (HMA) phenotype of MAF1 paralogs and

provided the first evidence that the ability to mediate HMA arose in a common ancestor of *T. gondii* and *H. hammondi* via neofunctionalization.

13. Freitas-Junior LH, Bottius E, Pirrit LA, Deitsch KW, Scheidig C, Guinet F, et al. Frequent ectopic recombination of virulence factor genes in telomeric chromosome clusters of *P. falciparum*. *Nature*. 2000;407(6807):1018–22. [PubMed: 11069183]
14. Pasternak ND, Dzikowski R. PfEMP1: an antigen that plays a key role in the pathogenicity and immune evasion of the malaria parasite *Plasmodium falciparum*. *Int J Biochem Cell Biol*. 2009;41(7):1463–6. [PubMed: 19150410]
15. Berriman M, Ghedin E, Hertz-Fowler C, Blandin G, Renauld H, Bartholomeu DC, et al. The genome of the African trypanosome *Trypanosoma brucei*. *Science*. 2005;309(5733):416–22. [PubMed: 16020726]
16. Mugnier MR, Cross GA, Papavasiliou FN. The in vivo dynamics of antigenic variation in *Trypanosoma brucei*. *Science*. 2015;347(6229):1470–3. [PubMed: 25814582]
17. Torgerson PR, Mastriacovo P. The global burden of congenital toxoplasmosis: a systematic review. *Bull World Health Organ*. 2013;91(7):501–8. [PubMed: 23825877]
18. Furtado JM, Smith JR, Belfort R, Gattley D, Jr, Winthrop KL. Toxoplasmosis: a global threat. *J Glob Infect Dis*. 2011;3(3):281–4. [PubMed: 21887062]
19. Lorenzi H, Khan A, Behnke MS, Namasivayam S, Swapna LS, Hadjithomas M, et al. Local admixture of amplified and diversified secreted pathogenesis determinants shapes mosaic *Toxoplasma gondii* genomes. *Nat Commun*. 2016;7:10147. [PubMed: 26738725] *[19]: Lorenzi et al., 2016 In this study the authors performed de novo sequencing and annotation of 63 *T. gondii* isolates from across the globe as well as a single isolate of *Hammondia hammondi*. Sequences and annotations were compared to *H. hammondi* and *Neospora caninum*, resulting in the conclusion that *T. gondii* could be distinguished from these other organisms based on genome-wide expansions of key gene families predicted to encode host- interacting secretory proteins.
20. Reese ML, Boothroyd JC. A conserved non-canonical motif in the pseudoactive site of the ROP5 pseudokinase domain mediates its effect on *Toxoplasma* virulence. *J Biol Chem*. 2011;286(33):29366–75. [PubMed: 21708941]
21. Wasmuth JD, Pszenny V, Haile S, Jansen EM, Gast AT, Sher A, et al. Integrated bioinformatic and targeted deletion analyses of the SRS gene superfamily identify SRS29C as a negative regulator of *Toxoplasma* virulence. *MBio*. 2012;3(6).
22. Reid AJ, Vermont SJ, Cotton JA, Harris D, Hill-Cawthorne GA, Konen-Waisman S, et al. Comparative genomics of the apicomplexan parasites *Toxoplasma gondii* and *Neospora caninum*: Coccidia differing in host range and transmission strategy. *PLoS Pathog*. 2012;8(3):e1002567. [PubMed: 22457617]
23. Walzer KA, Adomako-Ankomah Y, Dam RA, Herrmann DC, Schares G, Dubey JP, et al. *Hammondia hammondi*, an avirulent relative of *Toxoplasma gondii*, has functional orthologs of known *T. gondii* virulence genes. *Proc Natl Acad Sci U S A*. 2013;110(18):7446–51. [PubMed: 23589877]
24. Gajria B, Bahl A, Brestelli J, Dommer J, Fischer S, Gao X, et al. ToxoDB: an integrated *Toxoplasma gondii* database resource. *Nucleic Acids Res*. 2008;36(Database issue):D553–6. [PubMed: 18003657]
25. Boothroyd JC, Dubremetz JF. Kiss and spit: the dual roles of *Toxoplasma* rhoptries. *Nat Rev Microbiol*. 2008;6(1):79–88. [PubMed: 18059289]
26. Blader IJ, Coleman BI, Chen CT, Gubbels MJ. Lytic Cycle of *Toxoplasma gondii*. 15 Years Later. *Annu Rev Microbiol*. 2015;69:463–85. [PubMed: 26332089]
27. Saeij JP, Boyle JP, Collier S, Taylor S, Sibley LD, Brooke-Powell ET, et al. Polymorphic secreted kinases are key virulence factors in toxoplasmosis. *Science*. 2006;314(5806):1780–3. [PubMed: 17170306]
28. Behnke MS, Khan A, Wootton JC, Dubey JP, Tang K, Sibley LD. Virulence differences in *Toxoplasma* mediated by amplification of a family of polymorphic pseudokinases. *Proc Natl Acad Sci U S A*. 2011;108(23):9631–6. [PubMed: 21586633]

29. Reese ML, Zeiner GM, Saeij JP, Boothroyd JC, Boyle JP. Polymorphic family of injected pseudokinases is paramount in *Toxoplasma* virulence. *Proc Natl Acad Sci U S A*. 2011;108(23):9625–30. [PubMed: 21436047]
30. Behnke MS, Fentress SJ, Mashayekhi M, Li LX, Taylor GA, Sibley LD. The polymorphic pseudokinase ROP5 controls virulence in *Toxoplasma gondii* by regulating the active kinase ROP18. *PLoS Pathog*. 2012;8(11):e1002992. [PubMed: 23144612]
31. Pernas L, Adomako-Ankomah Y, Shastri AJ, Ewald SE, Treeck M, Boyle JP, et al. *Toxoplasma* effector MAF1 mediates recruitment of host mitochondria and impacts the host response. *PLoS Biol*. 2014;12(4):e1001845. [PubMed: 24781109]
32. Kim SK, Karasov A, Boothroyd JC. Bradyzoite-specific surface antigen SRS9 plays a role in maintaining *Toxoplasma gondii* persistence in the brain and in host control of parasite replication in the intestine. *Infect Immun*. 2007;75(4):1626–34. [PubMed: 17261600]
33. Hu K, Johnson J, Florens L, Fraunholz M, Suravajjala S, DiLullo C, et al. Cytoskeletal components of an invasion machine--the apical complex of *Toxoplasma gondii*. *PLoS Pathog*. 2006;2(2):e13. [PubMed: 16518471]
34. Gould SB, Tham WH, Cowman AF, McFadden GI, Waller RF. Alveolins, a new family of cortical proteins that define the protist infrakingdom Alveolata. *Mol Biol Evol*. 2008;25(6):1219–30. [PubMed: 18359944]
35. Olias P, Sibley LD. Functional Analysis of the Role of *Toxoplasma gondii* Nucleoside Triphosphate Hydrolases I and II in Acute Mouse Virulence and Immune Suppression. *Infect Immun*. 2016;84(7):1994–2001. [PubMed: 27091930]
36. Taylor S, Barragan A, Su C, Fux B, Fentress SJ, Tang K, et al. A secreted serine-threonine kinase determines virulence in the eukaryotic pathogen *Toxoplasma gondii*. *Science*. 2006;314(5806):1776–80. [PubMed: 17170305]
37. Behnke MS, Khan A, Lauron EJ, Jimah JR, Wang Q, Tolia NH, et al. Rhoptry Proteins ROP5 and ROP18 Are Major Murine Virulence Factors in Genetically Divergent South American Strains of *Toxoplasma gondii*. *PLoS Genet*. 2015;11(8):e1005434. [PubMed: 26291965]
38. Jung C, Lee CY, Grigg ME. The SRS superfamily of *Toxoplasma* surface proteins. *Int J Parasitol*. 2004;34(3):285–96. [PubMed: 15003490]
39. Crawford J, Lamb E, Wasmuth J, Grujic O, Grigg ME, Boulanger MJ. Structural and functional characterization of SporoSAG: a SAG2-related surface antigen from *Toxoplasma gondii*. *J Biol Chem*. 2010;285(16):12063–70. [PubMed: 20164173]
40. Jones TC, Yeh S, Hirsch J. The Interaction Between *Toxoplasma Gondii* and Mammalian Cells. *J Exp Med*. 1971;136(5):1157–72.
41. Sinai AP, Webster P, Joiner KA. Association of host cell endoplasmic reticulum and mitochondria with the *Toxoplasma gondii* parasitophorous vacuole membrane: a high affinity interaction. *J Cell Sci*. 1997;110 (Pt 17):2117–28. [PubMed: 9378762]
42. Pernas L, Boothroyd JC. Association of host mitochondria with the parasitophorous vacuole during *Toxoplasma* infection is not dependent on rhoptry proteins ROP2/8. *Int J Parasitol*. 2010;40(12):1367–71. [PubMed: 20637758]

Figure 1:

Our model for the stepwise evolution of the *MAF1* locus in *T. gondii* and its close relatives, *N. caninum* and *H. hammondi*. The MAF1a and MAF1b crystal structures ([11]) are in blue and orange, respectively. MAF1 maintains a vestigial ADPr binding macrodomain while duplication, expansion and GEDD events give rise to the neofunctionalized version of MAF1b driven at least in part by mutations of highly charged residues in the MAF1 C-terminus. Each significant change to the locus is marked and numbers correspond to MAF1 protein products in the key. MAF1b paralogs undergo subsequent diversification events following step 3 leading to the diverse shades of orange/yellow in the final *T. gondii* *MAF1* locus.

Figure 2:

a) Distribution of sequence gaps (black ticks) across all 14 *T. gondii* chromosomes and their co-localization with known tandem gene expansions (red blocks). Sequence gaps were identified by stretches of “N”s in version 30 of the *T. gondii* ME49 genome assembly which was downloaded from ToxoDB.org. Known tandem gene expansions were taken from the supplementary table in Adomako-Ankomah et al., mBio, 2014. b) Histogram of the location of sequence gaps in the *T. gondii* ME49 genome assembly (v30; www.toxodb.org) with respect to the nearest chromosome end, restricted to those on the 5 longest chromosomes (XII, XI, X, IX and VIII).