Published in final edited form as:

Trends Parasitol. 2011 August ; 27(8): 345–354. doi:10.1016/j.pt.2011.03.006.

Genome cartography: charting the apicomplexan genome

Jessica C. Kissinger^{1,2,3} and Jeremy DeBarry¹

¹Center for Tropical and Emerging Global Diseases, University of Georgia, Athens GA USA 30602

²Department of Genetics, University of Georgia, Athens GA USA 30602

³Institute of Bioinformatics, University of Georgia, Athens GA USA 30602

Abstract

Genes reside in particular genomic contexts that can be mapped at many levels. Historically, genetic maps were used primarily to locate genes. Recent technological advances in genome sequence determination have made the analysis and comparison of whole genomes possible and increasingly tractable. If we shift our focus from gene content (the inventory of genes contained within a genome), to genome composition and organization, what do we see? This review examines what has been learned about the evolution of the apicomplexan genome and the significance and impact of genomic location on our understanding of the eukaryotic genome and parasite biology.

The genomic feature landscape

Genomes are more than just vehicles for the transmission of genes; they have characteristic features beyond gene content. Eukaryotic genomes are organized as linear chromosomes that vary in size and number. Each chromosome has consistent regions, namely a centromere (to facilitate spindle attachment and movement of chromosomes during meiosis and mitosis), telomeres (to protect the ends of chromosomes) sub-telomeric and pericentromeric regions (located at the ends of chromosomes adjacent to telomeres and adjacent to centromeres, respectively) and origins of replication (where DNA replication is initiated) among others. A number of other features (genes, enhancers, repetitive elements) are located along the length of the chromosome at varying densities and with differing distributions. Each chromosome is condensed to chromatin by being wrapped around a core of histone proteins creating nucleosomes. Nucleosomes are dynamic and can ‘slide’ up and down regions of the chromosome to facilitate transcription according to the state of histone modifications in the nucleosome core. Nucleosome positioning is not uniform, and different regions of the genome are bound to differing extents giving rise to different states of chromatin such as euchromatin and heterochromatin.

In addition to this fundamental architecture, genomes can have regions in which the DNA is chemically modified, regions that are subject to higher rates of recombination and regions that appear to evolve more quickly or slowly than other regions [1, 2]. It is currently

© 2011 Elsevier Ltd. All rights reserved.

Corresponding Author: Jessica C. Kissinger, Paul Coverdell Center, Rm 370, 500 D.W. Brooks Drive, University of Georgia, jkissing@uga.edu, Phone: +1-706-542-6562.

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.

possible to look across the phylum at the entire apicomplexan genomic landscape and learn how it is organized and how it has evolved. We will do this by examining the distribution and patterns of genomic features.

Apicomplexan genomes: The lay of the land

Apicomplexan parasites are unicellular protists that are responsible for many significant diseases of humans and animals including malaria, the AIDS-related diseases toxoplasmosis and cryptosporidiosis, theileriosis in cattle and eimeriosis in chickens, among others. As a result of their medical and veterinary importance, a large number of genome sequences have been generated for members of this phylum

[3-11], http://gsc.jcvi.org/projects/msc/toxoplasma_gondii/index.shtml and <http://www.sanger.ac.uk/Projects/Pathogens/> (Figure 1 inset).

A dozen apicomplexan genome sequences have provided a wealth of information and insight into the biology and evolution of their genomes. The data reveal relatively small genome sizes (Figure 1) with significantly reduced numbers of protein-encoding genes ranging from a low of 3,671 in *Babesia bovis* to a high of ~8,000 in *Toxoplasma gondii* [3-5, 7, 9, 10, 12]. For comparison, the human, *Drosophila*, *Caenorhabditis* and *Saccharomyces* genomes contain approximately 20,000 to 25,000, 13,600, 19,000, and 6,000 genes, respectively [13, 14]. Introns are present in all apicomplexan organisms, but their number and size vary across the phylum with a low (both in terms of size and number) in *Cryptosporidium* species and a high in *Toxoplasma*.

The genomes are organized as 4 to 14 linear chromosomes each with a centromere, subtelomeric regions, and telomeres. Not all centromeres have been annotated or functionally identified in all genomes. The position of the centromere along the chromosome is variable. In *B. bovis* three of four centromeres are in an acrocentric location at one end of the chromosome and one is submetacentric [7]. In *Plasmodium*, centromeres display an extreme A/T nucleotide bias [5, 15, 16]. In *T. gondii*, centromeres exhibit no nucleotide bias, have distinct epigenetic signatures and appear to remain associated with centrocones beyond mitosis [17]. Extensive, gene-poor heterochromatic regions do not appear to exist. However, smaller, transcriptionally inactive 'territories' or regions of chromosomes have been observed in the apicomplexan species where this has been examined [17-23].

There is some variation with respect to telomere repeat sequences. The vertebrate consensus sequence is 5'-GGGTTA-3' while *Plasmodium* is 5'-GGGTTYA-3', where Y = T/C [24]. *Cryptosporidium* is 5'-GGTTTA-3' [25], and *Eimeria* and *Toxoplasma* are 5'-GGGTTTA-3' [26, 27]. The piroplasm species, *Theileria* and *Babesia*, have variable telomere sequences of the form 5'-G₂₋₃ T₃₋₄ A-3' [6, 7].

Subtelomeric regions are located at the ends of each chromosome adjacent to the telomeres. Subtelomeric regions vary within and between organisms and can contain dozens of genes. In many apicomplexan organisms, most notably *Plasmodium* and *Theileria*, these regions contain tandem arrays of protein-encoding genes, often surface antigens like the *var* genes in *P. falciparum* which encode the erythrocyte membrane protein, PfEMP1 [5, 6, 9].

Most apicomplexan genomes display a fairly even distribution of genes across their chromosomes; there is, however, one exception to this rule, and this is *Eimeria tenella*. The one fully sequenced and annotated chromosome available for this organism displays an extremely segmented structure of 'feature-rich' and 'feature-poor' regions across the full length of the chromosome [28].

A feature of the genomes of parasitic organisms anywhere on the tree of life is gene loss relative to the nearest free-living organism [29-32], and the Apicomplexa are no exception. Since the host can supply many nutrients, parasites can tolerate the loss of significant numbers of genes. Most genes that are lost are related to the synthesis of particular amino acids, nucleotides or nutrients that can be obtained from the host [3-11].

Many apicomplexan genomes are AT-rich with *P. falciparum* being the most extreme at 80% AT [5]. *P. falciparum* also contains many low-complexity insertions in protein-encoding genes. These phenomena likely result from multiple mechanisms including slippage during DNA replication [33] and the ways in which double strand breaks are repaired [34]. The recombination rate is high in the Apicomplexa, so there are many potential opportunities to introduce change into the genome. In *P. falciparum* 1cM = ~17 kb [35] and in *T. gondii* it is ~104 kb [36]. It is also possible that loss of particular DNA repair and recombination enzymes may also contribute to nucleotide biases [37] as is seen in bacteria, where this phenomenon is better studied [38]. The loss of repair enzymes can also affect other properties of genomes. For example, thus far, only the *T. gondii* genome is annotated as containing the required enzymes for the non-homologous end-joining (NHEJ) double-strand break repair pathway. Knockout of a single gene involved in this pathway has led to laboratory strains with an increased efficiency in homologous recombination [39]. Rampant gene loss can also radically alter the genomic landscapes of genomes [12]. For example, the apicomplexan mitochondrial genome is so reduced that it is only ~6 kb in size, contains three protein encoding genes and has become linearized [40]. The loss of chromosomal regions, containing genes and/or other features, can have dynamic consequences on the spacing and distribution of the remaining genomic features including, genes, enhancers and repeat sequences. This could affect the three-dimensional structure of chromosomes within the nucleus (see below) and the state of chromatin remodeling. The wholesale loss of genes also has profound consequences on our ability to detect synteny between species (see below) and greatly affects the genomic landscape.

Finally, intergenic sequences must be considered as a feature of the genomic landscape. Due to the streamlined nature of apicomplexan genomes, intergenic spaces tend to be very short; at times only a few hundred nucleotides are present between the stop codon of one protein-encoding gene and the start codon of the next [10-12]. As a result, fewer repetitive sequences are found in the Apicomplexa relative to model organisms [3-11]. In particular, there is a marked paucity of transposable elements in the apicomplexan organisms studied thus far, and none have been found in any of the completed genome sequences in Figure 1, [3-11]. Transposable elements (TEs) are found in the apicomplexan sister groups, the dinoflagellates and the ciliates, and nearly everywhere else. Thus, they have most likely been removed or lost via the same processes that have led to genome streamlining. It is increasingly clear that TEs affect genome evolution and biology at nearly every level, making them integral features of genomic landscapes that can drive genome restructuring through their transpositional or recombinatory activities [41-43]. Taken together, these phenomena paint the picture of a group of organisms that is radically different from 'model' eukaryotes.

Orthologous genes: same feature, different landscape

How many genes are shared by all apicomplexan genomes sequenced thus far? The answer is surprisingly few. There are only about 1,000 orthologous genes shared by all sequenced apicomplexan species [44, 45]. Depending upon the genome size, this number represents only 12 to 34% of the protein-encoding genes. The remainder of the genes identified in these genomes are specific to particular lineages. To a great extent, the gene repertoire dictates the environments an organism is able to exploit. For parasitic species, this is particularly

important as the range of hosts can quickly become restricted if the genes necessary for a particular niche are lost.

When a gene is restricted to a particular lineage or species, the question arises as to whether the gene was acquired, created *de novo*, or inherited from an ancestral genome, but subsequently lost by all other species. The difference between these hypotheses is quite significant. If all lineage-specific genes were 'new' then 60 to 80% of the gene content in extant apicomplexans would have to have been acquired or created *de novo*. This explanation is actually more parsimonious than positing independent losses in all other lineages; however, this does not appear to be what happened. Gene loss appears to be the strongest evolutionary force at work in the Apicomplexa and other parasite genomes [5, 12, 46]. An analysis of genes that are shared by all apicomplexans, and lineage- or species-specific reveals that most shared genes have orthologs outside of the Apicomplexa, suggesting they are ancestral or acquired. However, many genes that are genus- or species-specific do not have orthologs outside of the Apicomplexa, suggesting that they have either been created *de novo* or have evolved beyond recognition [45]. Thus, when we compare genomes and look at the genes that are lineage-specific, it is possible that these genes have been retained by chance, or it is possible, even probable, that they are essential to the parasite and cannot be lost.

If orthologous and lineage-specific genes are mapped to the genome and their distribution observed, several interesting patterns emerge. The different classes of genes (orthologous and species-specific) have very different distribution patterns in some species. In *Plasmodium*, orthologous genes are randomly distributed along the chromosome whereas species-specific multi-copy genes are greatly enriched at chromosome ends. The same pattern was not observed as strongly in *Theileria* [44] (Figure 2), where the multi-copy genes found at chromosome ends are often genus-specific rather than species-specific.

Both the creation/acquisition and distribution of species-specific and multi-copy genes can affect genome architecture. Apicomplexan genome sequence data have shown that there are many species-specific and multi-copy gene families present in these species, and that these genes are often integral to host-species interactions [3-11, 16, 44, 47]. Given this apparent reliance on 'new' genes, the small sizes of the genomes are striking. Overall, it appears that the removal of DNA is outpacing its generation [12, 48]. It should be noted that species-specific genes can be difficult to annotate and current numbers should be considered lower limits. This also places the absence of TEs in sharp contrast. As noted above, TEs are sources of evolutionary innovation that have been shown to be involved in the generation of novel genetic material. By identifying the distribution of these types of genes it may be possible to locate hotspots for the generation of new genes and implicate a mechanism for their formation.

Acquired genes: feature transplants

The apicomplexan nuclear genome has been significantly affected by genes acquired via intracellular and lateral gene transfers (Box 1) [49-54]. Genes acquired by intracellular transfer have come from many distinct sources, the algal nuclear genome, the algal plastid genome (now the apicoplast) and from the mitochondrial genome. Several programs have been written to identify genes that are currently present in the apicomplexan nuclear genome that encode products targeted to the apicoplast and mitochondrial organelles [55, 56]. If the positions of these acquired genes are plotted on the nuclear genome landscape we see an apparently random distribution (Figure 3(a)). There is, however, a noted absence of insertions at the extreme ends of chromosomes for plastid-targeted genes, and an even stronger absence from chromosomal ends in the case of mitochondrial-targeted genes in the

malaria parasite, *P. falciparum*. Most of these genes were likely transferred to the nuclear genome a very long time ago, thus the extant distribution of genes is likely a poor indicator of their original positions. The one pattern that remains is the observation that there is something special about chromosome ends, in particular, subtelomeric regions that are either rich with, or devoid of, particular features.

The situation for genes acquired from other organisms (primarily bacterial sources) show a similar but slightly different pattern. The best-studied apicomplexan parasite for lateral gene transfer is *Cryptosporidium parvum* [50]. Acquired genes have been plotted in Figure 3(b). The distribution appears to be random; however, there are several instances of acquired genes that are positioned quite close together. In one instance the genes are bacterial paralogs, suggesting they were acquired together on a single piece of DNA. The sample size is too small to draw any conclusions.

Synteny: landscape patterns

Synteny (used here to indicate synteny and collinearity) can be used to discover the conservation of genome structure between species and to investigate trends and patterns of changes in genome architecture that have shaped genomic landscapes. Synteny studies rely on the identification of shared markers (usually protein-encoding orthologs) in different genomes. Once identified, the primary interest is in the presence or absence of orthologous genes and their chromosomal distributions relative to a reference sequence. Regions of synteny can be thought of as 'blocks' separated by non-syntenic areas. By determining the degree of syntenic conservation, it is possible to discover the extent of divergence that has occurred since the species last shared a common ancestor by examining the numbers and types of genomic rearrangements that have occurred.

Investigations of regions located between syntenic blocks (regions of rearrangement) can yield information on how genomes are changing. Synteny can be disrupted in a variety of ways. Orthologs can diverge sufficiently that they cannot be identified, thus interfering with detection. Insertions and deletions of DNA can alter the genomic landscape, respectively, introducing and removing sequences between syntenic genes or altering the genes themselves. Most synteny detection methods can account for the presence of intervening genes between syntenic markers, skipping over non-syntenic genes that are the result of insertions. Chromosomal rearrangements, (translocations, deletions and inversions) can disrupt synteny via the large-scale elimination or redistribution of markers. Given enough information, it is possible to discover the types of changes that have disrupted synteny and are shaping genomes.

Comparisons of genomes can be visualized by aligning linear representations of chromosomes vertically, one below the other as in Figure 4 or by drawing lines that compare syntenic regions between chromosomes as in Figure 5. Synteny studies in the Apicomplexa reveal that there is significant synteny between the genomes of organisms in the same genus (Figures 5 and 6) [6, 7, 16, 45, 47]. The methods used for detection across the phylum shown in Figure 6 did not find any synteny conserved across the entire phylum (note the lack of lines that cross the circle in Figure 6). This indicates that genome rearrangement has been sufficiently extensive to remove any three genes on the same chromosome and in the same orientation since these species last shared a common ancestor. Limited synteny, however, is present at larger distances, e.g. between *Plasmodium* and *Theileria/Babesia* (note the lines connecting short syntenic blocks of gene features that cut across the circle between the genera in Figure 6). This level of rearrangement is more extensive than any observed in the vertebrates or fungi and their genomes contain transposable elements capable of causing insertions, deletions and movement of segments of DNA.

Genome rearrangements are one of the most easily observed indicators of genome evolution. Within the Apicomplexa we now see that rapid changes other than gene gains and losses are taking place. The entire landscape appears to be in a state of flux relative to the much more static genomes of vertebrates, starting with the number of chromosomes in each species (Figure 1) and continuing to the localization of multi-gene families. Thus, identification of the distribution of features across the landscape takes on an even greater significance when one considers the potential regulatory consequences of changes in location.

Gene expression, the histone code and nuclear architecture: a global perspective

Similar to other eukaryotes, gene expression in the Apicomplexa is influenced not only by *cis*-acting promoter elements, but also by the chromatin structure (acetylation or methylation of histone proteins in nucleosomes) [57-59]. Several factors appear to influence the position and spacing of nucleosomes along a chromosome and we do not yet fully understand them. However, the position and spacing of nucleosomes effectively create ‘territories’ along chromosomes with different properties and states of activity. Second, regions of chromosomes that are transcriptionally active tend to be co-localized and located at particular ‘zones’ within the nucleus, often near the nuclear envelope. Epigenetic regulatory proteins are localized in distinct nuclear sub-compartments in *Plasmodium* [60]. Active regions of the same chromosome and regions of different chromosomes can colocalize through chromosome folding and long-range interactions [61]. These active zones can be visualized microscopically, and the active chromosome regions within them can be detected with techniques like chromosome conformation capture (3C) or variants thereof [62].

These potential mechanisms of gene expression control are of special interest in the Apicomplexa, where it has been shown that there is a general paucity of molecular factors for transcriptional control in *Plasmodium* species. Several studies have contributed to our understanding of the relationship between epigenetic modifications and their locations and transcription. Inhibition of histone deacetylase activity was shown to deregulate transcription in *P. falciparum* [63]. Mapping of genome-wide nucleosome occupancy in *P. falciparum* has shown that chromatin architecture changes radically over the course of the lifecycle [64, 65]. In model eukaryotes the norm is targeted chromatin reorganization. Evidence indicates that these changes are the main factors behind transcription control.

Within the Apicomplexa many of the enzymes involved in histone modification have been identified, and ChIP-chip and ChIP-seq studies to identify chromosomal regions with particular histone modifications have been performed for several species. Zones of transcriptional activity have also been visualized in *Plasmodium* and *Toxoplasma*, and in a few cases some of the chromosomal regions that are associated with these zones have been identified, most notably subtelomeric regions containing *var* genes in *Plasmodium* [57, 58, 60, 66].

The picture that is emerging from all of these studies is one in which particular chromosomal ‘environments’ are created based upon the positioning of nucleosomes and the modification state of their histones combined with a nuclear architecture in which particular chromosomal regions are colocalized [67-69]. So, how does this relate to synteny and the evolution of the apicomplexan genome? Observations have emerged recently that associate changes in the transcriptional profile of particular genes with changes in their genome location [70]. This has been seen in other species. It is entirely plausible that genes, when moved from one region of a chromosome to another region or to a different chromosome, will find themselves in a different chromosomal environment with different nucleosome positioning and hence have altered transcription.

Conclusions and outstanding questions

In this age of genomics, where the focus has been on determining gene inventories, how important is genomic context to understanding gene regulation, gene evolution and ultimately the biology of the organism? After all, most eukaryotes, including the Apicomplexa, do not have polycistronic transcription. So, does it matter which chromosome or which particular segment of a chromosome contains a particular gene or sequence feature? A survey of the apicomplexan genome landscape has yielded several possible answers to these questions.

One dramatic observation is that the apicomplexan genome is highly dynamic with significant genome rearrangement. Why? The rearrangements were created by chance, but have they become fixed by stochastic processes or is there an element of selection? Could it be that altered gene transcription resulting from a new genic environment following a rearrangement creates a favorable selective feature? This is a testable hypothesis. As more expression data are gathered from closely related species, a comparative analysis of the expression profiles with the synteny profiles can associate changes in expression with breaks in synteny.

When we examine the patterns of gene distributions along chromosomes it is also clear that the patterns differ by species. Many species display a pattern in which lineage-specific genes exist as multi-gene families and these genes are enhanced in sub-telomeric regions. What is so special about the biology of these chromosomal regions that makes them amenable to gene duplications (either their generation, their permanence or both) and yet so refractory to acquired gene insertions? More data are needed.

Advances in technology are facilitating our data acquisition at an unprecedented rate. However, all sequences are not created equal. For the genome cartographer the lists of mountains and rivers, genes and repetitive elements are useless without a landscape and topology upon which to map them. Short-read sequence technologies have facilitated the rapid acquisition of gene-content information but in the absence of complete, fully assembled genome sequences a problem arises. In organisms that are not amenable to genetic crosses, or for whom we are lacking a karyotype, we will be missing crucial information that will help us to understand the evolution and biology of the genome.

Acknowledgments

The authors would like to thank David Morrison, Rolf Enzeroth and Boris Striepen for the electron micrograph of *Sarcocystis neurona* used in the cartography image, and Daniel Skelton for assistance with artwork. This work was supported by NIH award R01 AI068908 to J.C.K.

Glossary Box

3C (Chromosome Conformation Capture) and 4C (3C on Chip)

A foundational, high-throughput technology for the identification and mapping of interactions between specific, distant chromosomal loci. Interacting chromosomal loci are fixed via cross-linking and the bound DNA processed via restriction digestion, ligation and quantitatively amplified via Polymerase Chain Reaction (PCR) or hybridization to an array chip.

Apicoplast

A non-photosynthetic plastid organelle present in most species of Apicomplexa. It is surrounded by multiple membranes and likely arose from the endosymbiosis of a red alga (Box 1, Figure I).

Acrocentric	This describes a chromosome where the p arm is present but is too small to be easily observed (see Centromere below).
Cartography	The study and craft of map making.
Centrocone	A specialized structure of the apicomplexan nucleus associated with the intra-nuclear mitotic spindle.
Centromere	An essential region of the eukaryotic chromosome that serves as the point of mitotic spindle attachment during cell division. It is usually located near the middle of the chromosome and is characteristically heterochromatic. By convention, the centromere is the separation point for the two 'arms' of the chromosome. These are referred to as the short 'p' arm and a longer 'q' arm.
Chromatin remodeling	The dynamic disruption and formation of nucleosome complexes that constitute chromatin. Chromatin remodeling involves the effective shifting of nucleosome cores along the DNA molecule. These changes can affect the state of the chromatin and subsequently alter spatial and temporal patterns of gene expression.
Collinearity	A specific form of synteny, where the same loci (usually genes) are present on corresponding chromosomes and in corresponding order in different genomes.
Euchromatin	A less condensed (relative to heterochromatin) form of chromatin that is characteristically gene rich and usually actively transcribed.
Heterochromatin	A highly condensed form of chromatin that usually contains repetitive DNA and is generally transcriptionally repressed. Centromeres and telomeres are typically heterochromatic.
Histone code	The collective term for the covalently bound modifications of histone tails. Histones bind DNA to form nucleosomes that are the basis of chromatin structure. Modification of histone tails plays an important role in the state of chromatin and thus transcriptional activity.
Intracellular Gene Transfer (IGT)	The transfer of genes from an endosymbiont or organelle to the nuclear genome of an organism. The products of these genes are often transferred back to the donor for use.
Lateral Gene Transfer (LGT)	The transfer of genes from one organism to another, also known as horizontal gene transfer (HGT). In parasitic organisms these genes are possible drug targets due to their potentially distant evolutionary relationship with the organismal genes.
Nucleosomes	Repeating units of eukaryotic genomic DNA packaged around cores of histone protein octamers. Nucleosomes reduce the binding of a variety of DNA-binding proteins like regulatory proteins including transcription factors, thus affecting processes like transcription, DNA repair and recombination.
Orthologous	Is a term used to describe the relationship between genes or features present in different species but sharing a common evolutionary descent, e.g. the genes encoding alpha-hemoglobin in humans and alpha-hemoglobin in chimpanzees are orthologous.

Submetacentric	A chromosome where the arms are of unequal length (see Centromere above).
Subtelomeric	Variably sized regions of eukaryotic chromosomes adjacent to and directly interior of telomeres. The genomic features that they contain usually define them. They are typically composed of highly polymorphic repetitive DNA sequences that are situated adjacent to gene-rich areas. In <i>Plasmodium</i> species, subtelomeric regions often contain genomic repeats and species-specific genes involved in host immune response evasion.
Synten	The presence of the same loci (usually genes) on corresponding chromosomes in different genomes.
Telomere	A specialized region of DNA composed of large numbers of short nucleotide repeats complexed with proteins, found at the ends of linear eukaryotic chromosomes. Telomeres protect the ends of chromosomes from degradation, acting as disposable buffers during chromosome replication.

References

1. Bonnivard E, Higuete D. Fluidity of eukaryotic genomes. C R Biol. 2009; 332:234–240. [PubMed: 19281954]
2. Eichler EE, Sankoff D. Structural dynamics of eukaryotic chromosome evolution. Science. 2003; 301:793–797. [PubMed: 12907789]
3. Carlton JM, et al. Comparative genomics of the neglected human malaria parasite *Plasmodium vivax*. Nature. 2008; 455:757–763. [PubMed: 18843361]
4. Carlton JM, et al. Genome sequence and comparative analysis of the model rodent malaria parasite *Plasmodium yoelii yoelii*. Nature. 2002; 419:512–519. [PubMed: 12368865]
5. Gardner MJ, et al. Genome sequence of the human malaria parasite *Plasmodium falciparum*. Nature. 2002; 419:498–511. [PubMed: 12368864]
6. Pain A, et al. Genome of the host-cell transforming parasite *Theileria annulata* compared with *T. parva*. Science. 2005; 309:131–133. [PubMed: 15994557]
7. Brayton KA, et al. Genome sequence of *Babesia bovis* and comparative analysis of apicomplexan hemoprotozoa. PLoS Pathog. 2007; 3:1401–1413. [PubMed: 17953480]
8. Pain A, et al. The genome of the simian and human malaria parasite *Plasmodium knowlesi*. Nature. 2008; 455:799–803. [PubMed: 18843368]
9. Gardner MJ, et al. Genome sequence of *Theileria parva*, a bovine pathogen that transforms lymphocytes. Science. 2005; 309:134–137. [PubMed: 15994558]
10. Abrahamsen MS, et al. Complete Genome Sequence of the Apicomplexan, *Cryptosporidium parvum*. Science. 2004; 304:441–445. [PubMed: 15044751]
11. Xu P, et al. The genome of *Cryptosporidium hominis*. Nature. 2004; 431:1107–1112. [PubMed: 15510150]
12. Keeling PJ. Reduction and compaction in the genome of the apicomplexan parasite *Cryptosporidium parvum*. Dev Cell. 2004; 6:614–616. [PubMed: 15130487]
13. International Human Genome Sequencing Consortium. Finishing the euchromatic sequence of the human genome. Nature. 2004; 431:931–945. [PubMed: 15496913]
14. Rubin GM, et al. Comparative genomics of the eukaryotes. Science. 2000; 287:2204–2215. [PubMed: 10731134]
15. Bowman S, et al. The complete nucleotide sequence of chromosome 3 of *Plasmodium falciparum*. Nature. 1999; 400:532–538. [PubMed: 10448855]

16. Kooij TW, et al. A *Plasmodium* Whole-Genome Synteny Map: Indels and Synteny Breakpoints as Foci for Species-Specific Genes. *PLoS Pathog.* 2005; 1:e44. [PubMed: 16389297]
17. Brooks CF, et al. *Toxoplasma gondii* sequesters centromeres to a specific nuclear region throughout the cell cycle. *Proc Natl Acad Sci U S A.* 2011; 108:3767–3772. [PubMed: 21321216]
18. Dzikowski R, Deitsch KW. Genetics of antigenic variation in *Plasmodium falciparum*. *Curr Genet.* 2009; 55:103–110. [PubMed: 19242694]
19. Voss TS, et al. A var gene promoter controls allelic exclusion of virulence genes in *Plasmodium falciparum* malaria. *Nature.* 2006; 439:1004–1008. [PubMed: 16382237]
20. Flueck C, et al. *Plasmodium falciparum* heterochromatin protein 1 marks genomic loci linked to phenotypic variation of exported virulence factors. *PLoS Pathog.* 2009; 5:e1000569. [PubMed: 19730695]
21. Tonkin CJ, et al. Sir2 paralogues cooperate to regulate virulence genes and antigenic variation in *Plasmodium falciparum*. *PLoS Biol.* 2009; 7:e84. [PubMed: 19402747]
22. Deitsch KW. A mark of silence in malaria parasites. *Cell Host Microbe.* 2009; 5:112–113. [PubMed: 19218082]
23. Epp C, et al. Chromatin associated sense and antisense noncoding RNAs are transcribed from the var gene family of virulence genes of the malaria parasite *Plasmodium falciparum*. *RNA.* 2009; 15:116–127. [PubMed: 19037012]
24. Scherf A, et al. *Plasmodium* telomeres: a pathogen's perspective. *Curr Opin Microbiol.* 2001; 4:409–414. [PubMed: 11495803]
25. Liu C, et al. Telomeric sequences of *Cryptosporidium parvum*. *Mol Biochem Parasitol.* 1998; 94:291–296. [PubMed: 9747979]
26. Shirley MW. The genome of *Eimeria* spp., with special reference to *Eimeria tenella*—a coccidium from the chicken. *Int J Parasitol.* 2000; 30:485–493. [PubMed: 10731571]
27. Gajria B, et al. ToxoDB: an integrated *Toxoplasma gondii* database resource. *Nucleic Acids Res.* 2008; 36:D553–556. [PubMed: 18003657]
28. Ling KH, et al. Sequencing and analysis of chromosome 1 of *Eimeria tenella* reveals a unique segmental organization. *Genome Res.* 2007; 17:311–319. [PubMed: 17284678]
29. Keeling PJ, Fast NM. Microsporidia: biology and evolution of highly reduced intracellular parasites. *Annu Rev Microbiol.* 2002; 56:93–116. [PubMed: 12142484]
30. Sakharkar KR, et al. Genome reduction in prokaryotic obligatory intracellular parasites of humans: a comparative analysis. *Int J Syst Evol Microbiol.* 2004; 54:1937–1941. [PubMed: 15545414]
31. Moran NA. Tracing the evolution of gene loss in obligate bacterial symbionts. *Curr Opin Microbiol.* 2003; 6:512–518. [PubMed: 14572545]
32. Moran NA, Plague GR. Genomic changes following host restriction in bacteria. *Curr Opin Genet Dev.* 2004; 14:627–633. [PubMed: 15531157]
33. Ellegren H. Microsatellites: simple sequences with complex evolution. *Nat Rev Genet.* 2004; 5:435–445. [PubMed: 15153996]
34. Zilversmit MM, et al. Low-complexity regions in *Plasmodium falciparum*: missing links in the evolution of an extreme genome. *Mol Biol Evol.* 2010; 27:2198–2209. [PubMed: 20427419]
35. Su X, et al. A genetic map and recombination parameters of the human malaria parasite *Plasmodium falciparum*. *Science.* 1999; 286:1351–1353. [PubMed: 10558988]
36. Khan A, et al. Composite genome map and recombination parameters derived from three archetypal lineages of *Toxoplasma gondii*. *Nucleic Acids Res.* 2005; 33:2980–2992. [PubMed: 15911631]
37. Ralph SA, et al. Evolutionary pressures on apicoplast transit peptides. *Molecular biology and evolution.* 2004; 21:2183–2194. [PubMed: 15317876]
38. Lind PA, Andersson DI. Whole-genome mutational biases in bacteria. *Proc Natl Acad Sci U S A.* 2008; 105:17878–17883. [PubMed: 19001264]
39. Fox BA, et al. Efficient gene replacements in *Toxoplasma gondii* strains deficient for nonhomologous end joining. *Eukaryot Cell.* 2009; 8:520–529. [PubMed: 19218423]
40. Hikosaka K, et al. Divergence of the mitochondrial genome structure in the apicomplexan parasites, *Babesia* and *Theileria*. *Mol Biol Evol.* 2010; 27:1107–1116. [PubMed: 20034997]

41. Bennetzen JL. Transposable element contributions to plant gene and genome evolution. *Plant Mol Biol.* 2000; 42:251–269. [PubMed: 10688140]
42. Bartolome C, et al. On the abundance and distribution of transposable elements in the genome of *Drosophila melanogaster*. *Mol Biol Evol.* 2002; 19:926–937. [PubMed: 12032249]
43. Bohne A, et al. Transposable elements as drivers of genomic and biological diversity in vertebrates. *Chromosome Res.* 2008; 16:203–215. [PubMed: 18293113]
44. Kuo CH, Kissinger JC. Consistent and contrasting properties of lineage-specific genes in the apicomplexan parasites *Plasmodium* and *Theileria*. *BMC Evol Biol.* 2008; 8:108. [PubMed: 18405380]
45. DeBarry J, Kissinger J. Jumbled Genomes: Missing Apicomplexan Synteny. *Molecular Biology and Evolution.* 2011; 28
46. Slamovits CH, et al. Genome compaction and stability in microsporidian intracellular parasites. *Curr Biol.* 2004; 14:891–896. [PubMed: 15186746]
47. Weir W, et al. Highly syntenic and yet divergent: A tale of two *Theilerias*. *Infection, Genetics and Evolution.* 2009; 9:453–461.
48. Mira A, et al. Deletional bias and the evolution of bacterial genomes. *Trends Genet.* 2001; 17:589–596. [PubMed: 11585665]
49. Zhu G, Keithly JS. Alpha-proteobacterial relationship of apicomplexan lactate and malate dehydrogenases. *J Eukaryot Microbiol.* 2002; 49:255–261. [PubMed: 12120991]
50. Huang J, et al. Phylogenomic evidence supports past endosymbiosis, intracellular and horizontal gene transfer in *Cryptosporidium parvum*. *Genome Biol.* 2004; 5:R88. [PubMed: 15535864]
51. Huang J, et al. A first glimpse into the pattern and scale of gene transfer in Apicomplexa. *Int J Parasitol.* 2004; 34:265–274. [PubMed: 15003488]
52. Huang, J.; Kissinger, J. Horizontal and intracellular gene transfer in the Apicomplexa: The scope and functional consequences. In: Katz, L.; Bhattacharya, D., editors. *Genome Evolution in Eukaryotic Microbes*. Oxford University Press; 2006. p. 256
53. Striepen B, et al. Gene transfer in the evolution of parasite nucleotide biosynthesis. *Proceedings of the National Academy of Sciences, USA.* 2004; 101:3154–3159.
54. Nagamune K, Sibley LD. Comparative genomic and phylogenetic analyses of calcium ATPases and calcium-regulated proteins in the apicomplexa. *Mol Biol Evol.* 2006; 23:1613–1627. [PubMed: 16751258]
55. Zuegge J, et al. Deciphering apicoplast targeting signals--feature extraction from nuclear-encoded precursors of *Plasmodium falciparum* apicoplast proteins. *Gene.* 2001; 280:19–26. [PubMed: 11738814]
56. Foth BJ, et al. Dissecting apicoplast targeting in the malaria parasite *Plasmodium falciparum*. *Science.* 2003; 299:705–708. [PubMed: 12560551]
57. Westenberger SJ, et al. Genome-wide nucleosome mapping of *Plasmodium falciparum* reveals histone-rich coding and histone-poor intergenic regions and chromatin remodeling of core and subtelomeric genes. *BMC Genomics.* 2009; 10:610. [PubMed: 20015349]
58. Gissot M, Kim K. How epigenomics contributes to the understanding of gene regulation in *Toxoplasma gondii*. *J Eukaryot Microbiol.* 2008; 55:476–480. [PubMed: 19120792]
59. Sullivan WJ Jr. et al. Histones and histone modifications in protozoan parasites. *Cell Microbiol.* 2006; 8:1850–1861. [PubMed: 17026479]
60. Volz J, et al. Potential epigenetic regulatory proteins localise to distinct nuclear sub-compartments in *Plasmodium falciparum*. *Int J Parasitol.* 2010; 40:109–121. [PubMed: 19765590]
61. Lieberman-Aiden E, et al. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science.* 2009; 326:289–293. [PubMed: 19815776]
62. Bartova E, et al. Histone modifications and nuclear architecture: a review. *J Histochem Cytochem.* 2008; 56:711–721. [PubMed: 18474937]
63. Chaal BK, et al. Histone deacetylases play a major role in the transcriptional regulation of the *Plasmodium falciparum* life cycle. *PLoS Pathog.* 6:e1000737. [PubMed: 20107518]
64. Ponts N, et al. Nucleosome landscape and control of transcription in the human malaria parasite. *Genome Res.* 2010; 20:228–238. [PubMed: 20054063]

65. Ponts N, et al. Nucleosome occupancy at transcription start sites in the human malaria parasite: A hard-wired evolution of virulence? *Infect Genet Evol.*
66. Gissot M, et al. Epigenomic modifications predict active promoters and gene structure in *Toxoplasma gondii*. *PLoS Pathog.* 2007; 3:e77. [PubMed: 17559302]
67. van Steensel B, Dekker J. Genomics tools for unraveling chromosome architecture. *Nat Biotechnol.* 28:1089–1095. [PubMed: 20944601]
68. Gondor A, Ohlsson R. Chromosome crosstalk in three dimensions. *Nature.* 2009; 461:212–217. [PubMed: 19741702]
69. Tsankov AM, et al. The role of nucleosome positioning in the evolution of gene regulation. *PLoS Biol.* 8:e1000414. [PubMed: 20625544]
70. Bozdech Z, et al. The transcriptome of *Plasmodium vivax* reveals divergence and diversity of transcriptional regulation in malaria parasites. *Proc Natl Acad Sci U S A.* 2008; 105:16290–16295. [PubMed: 18852452]
71. Kohler S, et al. A plastid of probable green algal origin in Apicomplexan parasites. *Science.* 1997; 275:1485–1489. [PubMed: 9045615]
72. Lau AO, et al. *Babesia bovis*: a comprehensive phylogenetic analysis of plastid-encoded genes supports green algal origin of apicoplasts. *Exp Parasitol.* 2009; 123:236–243. [PubMed: 19646439]
73. Fast NM, et al. Nuclear-encoded, plastid-targeted genes suggest a single common origin for apicomplexan and dinoflagellate plastids. *Mol Biol Evol.* 2001; 18:418–426. [PubMed: 11230543]
74. Moore RB, et al. A photosynthetic alveolate closely related to apicomplexan parasites-original article. *Nature.* 2008; 451:959–963. [PubMed: 18288187]
75. Janouskovec, j., et al. A common red algal origin of the apicomplexan, dinoflagellate, and heterokont plastids. *Proc Natl Acad Sci U S A.* 2009; 107:10949–10954. [PubMed: 20534454]
76. Lim L, et al. New proteins in the apicoplast membranes: time to rethink apicoplast protein targeting. *Trends Parasitol.* 2009; 25:197–200. [PubMed: 19346163]
77. Waller RF, et al. Nuclear-encoded proteins target to the plastid in *Toxoplasma gondii* and *Plasmodium falciparum*. *Proc Natl Acad Sci U S A.* 1998; 95:12352–12357. [PubMed: 9770490]
78. Martin W. Gene transfer from organelles to the nucleus: frequent and in big chunks. *Proc Natl Acad Sci U S A.* 2003; 100:8612–8614. [PubMed: 12861078]
79. Martin W, et al. Gene transfer to the nucleus and the evolution of chloroplasts. *Nature.* 1998; 393:162–165. [PubMed: 11560168]
80. Sequence and comparative analysis of the chicken genome provide unique perspectives on vertebrate evolution. *Nature.* 2004; 432:695–716. [PubMed: 15592404]
81. Abubucker S, et al. The canine hookworm genome: analysis and classification of *Ancylostoma caninum* survey sequences. *Mol Biochem Parasitol.* 2008; 157:187–192. [PubMed: 18082904]
82. Adam RD. The *Giardia lamblia* genome. *Int J Parasitol.* 2000; 30:475–484. [PubMed: 10731570]
83. Arumuganathan K, Earle E. Nuclear DNA content of some important plant species. *Plant Molecular Biology Reporter.* 1991; 9:208–218.
84. Berriman M, et al. The genome of the African trypanosome *Trypanosoma brucei*. *Science.* 2005; 309:416–422. [PubMed: 16020726]
85. Berriman M, et al. The genome of the blood fluke *Schistosoma mansoni*. *Nature.* 2009; 460:352–358. [PubMed: 19606141]
86. Carlton JM, et al. Draft genome sequence of the sexually transmitted pathogen *Trichomonas vaginalis*. *Science.* 2007; 315:207–212. [PubMed: 17218520]
87. Corradi N, et al. The complete sequence of the smallest known nuclear genome from the microsporidian *Encephalitozoon intestinalis*. *Nat Commun.* 1:77. [PubMed: 20865802]
88. Ivens AC, et al. The genome of the kinetoplastid parasite, *Leishmania major*. *Science.* 2005; 309:436–442. [PubMed: 16020728]
89. Loftus B, et al. The genome of the protist parasite *Entamoeba histolytica*. *Nature.* 2005; 433:865–868. [PubMed: 15729342]
90. Mitreva M, Jasmer DP. Biology and genome of *Trichinella spiralis*. *WormBook.* 2006:1–21. [PubMed: 18050431]

91. Scott AL, Ghedin E. The genome of *Brugia malayi* - all worms are not created equal. *Parasitol Int.* 2009; 58:6–11. [PubMed: 18952001]
92. Venter JC, et al. The sequence of the human genome. *Science.* 2001; 291:1304–1351. [PubMed: 11181995]
93. Aurrecochea C, et al. PlasmoDB: a functional genomic database for malaria parasites. *Nucleic Acids Res.* 2009; 37:D539–543. [PubMed: 18957442]

Box 1. Evolution of extant apicomplexan species and their organelles

Apicomplexan parasites are unicellular eukaryotes that have arisen as the result of a secondary endosymbiosis in which a eukaryotic cell endosymbiosed an alga (Figure I). There has been some debate over the years about the type of alga that was endosymbiosed [71-73], but the current evidence points towards a red algal ancestor [74, 75]. Since algae are photosynthetic as a result of the primary endosymbiotic event that gave rise to photosynthetic eukaryotes, the Apicomplexa were able to keep this organelle, now termed the apicoplast. The apicoplast is surrounded by the original plastid membranes and two additional membranes believed to be the outer algal membrane and recipient host membrane [76]. The genes involved in photosynthesis and many other processes have been lost. Many others have been transferred to the nuclear genome, and their products are targeted back to the organelle where they function [77]. This rare and unusual evolutionary history has provided many possible sources of genetic material that has been transferred to the sole remaining nuclear genome. Gene transfer from organelles (mitochondria and chloroplasts) to nuclear genomes is not a new phenomenon; it has occurred extensively in most eukaryotic cells that contain them [78, 79].

Figure I Schematic of evolutionary events affecting the apicomplexan nuclear genome

(a) Primary endosymbiosis of a cyanobacterium (green) by a eukaryotic cell to create a photosynthetic eukaryotic cell; **(b)** Transfer of genes from the chloroplast (green) and the mitochondria (orange) to the nuclear genome (black) in an algal cell; **(c)** Secondary endosymbiosis of a red alga by another eukaryotic cell. Continued gene transfer from the algal nucleus and chloroplast to the new 'host' nuclear genome; **(d)** Loss of the algal nucleus to give rise to modern apicomplexan cell; **(e)** Lateral gene transfer (LGT) from external sources (red) is occurring throughout evolution of the Apicomplexa. Gene transfer continues from organelles to the nuclear genome. This figure is reproduced in part from [52].

Figure 1. Genome sizes and relationships of select parasitic eukaryotes

Approximate genome sizes are shown in Mb. The upper scale is for select model organisms and the lower scale is for selected parasites. Apicomplexan genomes are shown in red. Parasites exhibit reduced genome sizes relative to other eukaryotes. The inset depicts the relationships of select apicomplexan parasites. Sources for the sizes of apicomplexan genome sizes are located in the text. Other eukaryotic model and parasite genome sizes are from their respective genome projects, papers and NCBI GenBank [80-92].

Figure 2. Distribution of species-specific, shared, and multicopy genes in *P. falciparum* and *T. annulata*

(a) The outer circle represents the chromosomes of *P. falciparum* and *T. annulata*. Ticks = 100 kb. Chromosome numbers are indicated following the species abbreviation. Inner circular tick tracks represent the position of species-specific (red), shared (black) and multicopy (blue) genes. Shared genes have at least one ortholog in the species shown in the inset of Figure 1, with the exception of *E. tenella*. (b) Enlarged view of the largest chromosomes from each species. Both species have an enrichment of multi-copy genes at chromosome ends. However, *P. falciparum* shows more species-specific genes than *T. annulata*. [44]. Genes from every species in the inset from Figure 1, except *E. tenella*, were used to identify species-specific and orthologous genes. Therefore the patterns observed are not the result of comparison between only *P. falciparum* and *T. annulata*. Abbreviations: Ta, *Theileria annulata*; Pf, *Plasmodium falciparum*. Data for both (a) and (b) are adapted from [44, 45].

Figure 3. Distribution of intracellularly and laterally transferred genes

The outermost colored circles represent the chromosomes of (a) *P. falciparum* putative intracellular gene transfers. The green tick track indicates the position of nuclear-encoded genes whose products are targeted to the apicoplast, and the red tick track indicates products targeted to the mitochondria. Gene products with putative subcellular localizations in the apicoplast or mitochondrion (based on targeting sequences or metabolic pathways restricted to these organelles) were identified using search tools at PlasmoDB [93] (b) *C. parvum* positions of laterally acquired genes [50]. Red ticks indicate the position of genes directly adjacent to each other. Ticks = 100 kb. Chromosome numbers are indicated following the species abbreviation. Abbreviations: Pf, *Plasmodium falciparum*; Cp, *Cryptosporidium parvum*.

Figure 4. Synteny and genome rearrangement

Horizontal lines represent sections of chromosomes from different genomes. Colored boxes represent synteny markers, with the same colors representing orthologous genes. Black boxes represent non-orthologous genes that are not used as markers. Chromosomes and genes are not drawn to scale. **(a)** The top sequence is an arbitrary reference sequence for the order of orthologous genes. Despite different numbers and distributions of non-orthologous genes, the six orthologous markers are conserved in a syntenic block with the comparator sequence. Sequences are drawn so that the corresponding markers align vertically. **(b)** Synteny (relative to the Reference in **(a)**) can be disrupted locally within chromosomes or at the level of the whole chromosome. Local changes can take the form of deletions (yellow 'x'), duplications (blue boxes outlined in heavy black) or insertions (intervening non-orthologous genes on all chromosome sections can be the result of insertion events). The chromosome section shown in red is the result of a translocation event that has fused two chromosomes, removing half of the genes in the syntenic block. This pattern could also be derived from the deletion or inversion of large chromosome segments.

Figure 5. Synteny across the *Piroplasmida*

The circle is a graphical representation of the annotated chromosomes and contigs in each piroplasmid genome (*Babesia* and *Theileria*). Each species' genome is labeled with the genus-species abbreviation. Unordered sections of chromosomes are labeled with '-'. Ticks = 100 kb. Spans that cross the interior of the circle connect syntenic regions identified with protein-coding ortholog markers. Spans that narrow in the middle are 'twisted' and represent inversions. Different colors represent different chromosomes within each species. Figure modified from [45]. Abbreviations: Bb, *Babesia bovis*; Tp, *Theileria parva*; Ta, *Theileria annulata*.

Figure 6. Synteny across the Apicomplexa

The circle is a graphical representation of the annotated chromosomes and contigs in each genome. Each species' genome is labeled with the genus-species abbreviation. Scaffolds/Contigs that are not assigned to chromosomes, but contain syntenic regions are shown in black. Tick marks represent 1 Mb. Spans that cross the interior of the circle connect syntenic regions. Spans that narrow in the middle are 'twisted' and represent inversions. Different colors represent different chromosomes within each species. Figure modified from [45]. Abbreviations: Pf, *Plasmodium falciparum*; Pv, *Plasmodium vivax*; Ta, *Theileria annulata*; Bb, *Babesia bovis*; Tg, *Toxoplasma gondii*; Nc, *Neospora caninum*; Cp, *Cryptosporidium parvum*; Cm, *Cryptosporidium muris*.