Jumbled Genomes: Missing Apicomplexan Synteny

Jeremy D. DeBarry*[1,2] and Jessica C. Kissinger[1,2,3]

[1]Center for Tropical and Emerging Global Diseases, University of Georgia

[2]Department of Genetics, University of Georgia

[3]Institute of Bioinformatics, University of Georgia

***Corresponding author:** E-mail: jdebarry@uga.edu.

Associate editor: Hervé Philippe

Research article

Abstract

Whole-genome comparisons provide insight into genome evolution by informing on gene repertoires, gene gains/losses, and genome organization. Most of our knowledge about eukaryotic genome evolution is derived from studies of multicellular model organisms. The eukaryotic phylum Apicomplexa contains obligate intracellular protist parasites responsible for a wide range of human and veterinary diseases (e.g., malaria, toxoplasmosis, and theileriosis). We have developed an *in silico* protein-encoding gene based pipeline to investigate synteny across 12 apicomplexan species from six genera. Genome rearrangement between lineages is extensive. Syntenic regions (conserved gene content and order) are rare between lineages and appear to be totally absent across the phylum, with no group of three genes found on the same chromosome and in the same order within 25 kb up- and downstream of any orthologous genes. Conserved synteny between major lineages is limited to small regions in *Plasmodium* and *Theileria*/*Babesia* species, and within these conserved regions, there are a number of proteins putatively targeted to organelles. The observed overall lack of synteny is surprising considering the divergence times and the apparent absence of transposable elements (TEs) within any of the species examined. TEs are ubiquitous in all other groups of eukaryotes studied to date and have been shown to be involved in genomic rearrangements. It appears that there are different criteria governing genome evolution within the Apicomplexa relative to other well-studied unicellular and multicellular eukaryotes.

Key words: eukaryotic genome evolution, gene loss, genome architecture, genome rearrangement, parasite, protist.

Introduction

Conservation of gene content and genome organization is usually correlated with divergence time, especially in eukaryotes. Synteny and collinearity (the conserved content and order of genetic loci, respectively, on the same chromosome, referred to here as synteny) are usually detectable and often prevalent among related eukaryotes. For example, synteny can be detected over 550 My between the chordate amphioxous and humans (Putnam et al. 2008). The phylum Apicomplexa is an exception to this trend, with an astonishing lack of synteny between different lineages, despite divergence times equal to or less than, those of other eukaryotes with a high degree of syntenic conservation (see below). This difference is perhaps less surprising considering that the vast majority of what is known about eukaryotic genome evolution has been learned from the focused study of a few, primarily multicellular, model organisms. However, the bulk of eukaryotic diversity is represented by unicelluar organisms (Baldauf 2003), and the dynamics of their genome architecture remain largely unexplored. The availability of whole-genome sequence data for many members of the protistan parasite phylum Apicomplexa offers a unique opportunity to investigate the genome-scale patterns and trends that have occurred with the evolution of parasitism.

The Apicomplexa

Malaria is the most notorious human disease caused by apicomplexan organisms. The phylum also contains the AIDS-related *Cryptosporidium* and *Toxoplasma* pathogens as well as several other pathogens of human and veterinary importance. Genomes in the Apicomplexa are extremely small (~8.5–63 Mb, fig. 1) relative to many sequenced eukaryotes (Carlton et al. 2002, 2008; Gardner et al. 2002, 2005; Abrahamsen et al. 2004; Brayton et al. 2007). They are characterized by gene loss (Kuo and Kissinger 2008), with only a few thousand protein-encoding genes per genome and both intracellular and lateral gene transfer (Zhu and Keithly 2002; Huang, Mullapudi, Lancto, et al. 2004; Huang, Mullapudi, Sicheritz-Ponten, and Kissinger 2004; Striepen et al. 2004; Huang and Kissinger 2006; Nagamune and Sibley 2006). The most striking example of gene loss to date is found in *Cryptosporidium parvum*, where all pathways for *de novo* nucleotide synthesis have been lost and nucleotide salvage pathways have been acquired (Striepen et al. 2004). This phenomenon of significant gene loss and lateral gene acquisition and their effect on shaping genome architecture cannot be studied in model eukaryotic organisms.

One of the most unanticipated discoveries in apicomplexan genomes is the near absence of transposable

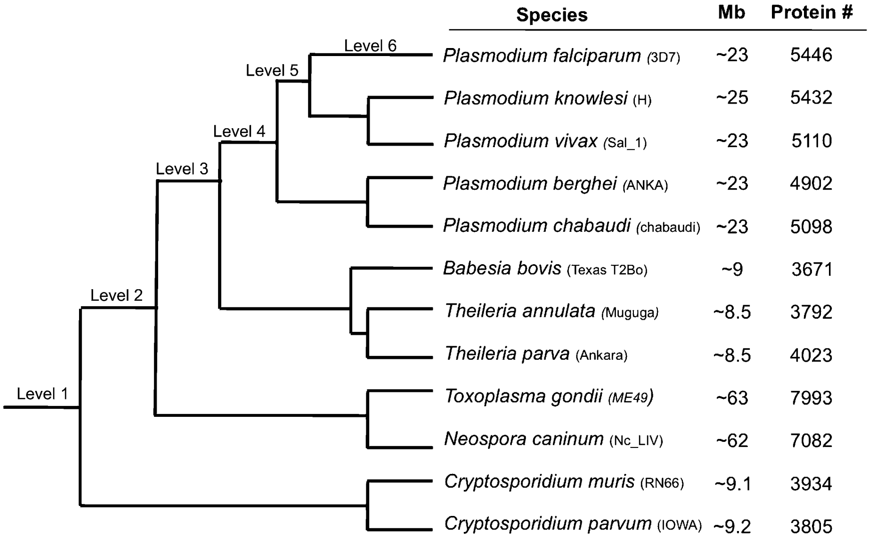
Species
Mb
Protein #
Level 1
Level 2
Level 3
Level 4
Level 5
Level 6
Plasmodium falciparum (3D7) ~23 5446
Plasmodium knowlesi (H) ~25 5432
Plasmodium vivax (Sal_1) ~23 5110
Plasmodium berghei (ANKA) ~23 4902
Plasmodium chabaudi (chabaudi) ~23 5098
Babesia bovis (Texas T2Bo) ~9 3671
Theileria annulata (Muguga) ~8.5 3792
Theileria parva (Ankara) ~8.5 4023
Toxoplasma gondii (ME49) ~63 7993
Neospora caninum (Nc_LIV) ~62 7082
Cryptosporidium muris (RN66) ~9.1 3934
Cryptosporidium parvum (IOWA) ~9.2 3805

Fig. 1. Species relationships and genome characteristics. A cladogram of investigated species with genome sizes, numbers of annotated protein-encoding genes, and numbers of chromosomes. Numbered labels on the cladogram indicate the different levels examined for ortholog cluster distribution. For example, level 1 contains ortholog clusters with members in all Apicomplexa. Level 5 contains ortholog clusters specific to *Plasmodium falciparum*, *P. knowlesi*, and *P. vivax* and not detected in the other apicomplexan species.

elements (TEs). TEs, defined by their ability to mobilize and replicate within a host, increasing their copy number, are a ubiquitous feature of all other investigated eukaryotic genomes. Despite genomic data for 12 members of the phylum and rigorous examination of several species with both similarity based and *de novo* methods (Barrie A, Cheng S, Kissinger J, Pritham E, personal communication), evidence of TEs remains sparse. Only a handful of putative apicomplexan TEs and associated protein domains have been reported (Durand et al. 2006; Templeton et al. 2009). Through their transpositional activities or their function as sites for ectopic recombination, TEs are major drivers of genome rearrangement (Bennetzen 2000; Bartolome et al. 2002; Kent et al. 2003; Hua-Van et al. 2005; Bohne et al. 2008). The loss of TEs from most members of the apicomplexan lineage leaves the agent behind their extensive loss of synteny an intriguing anomaly in eukaryotic genome evolution.

Genome Rearrangement: Eukaryotic Examples

Comparisons of genome architecture are designed to discover differences and similarities, and the underlying causes and importance of each. Genome rearrangements are a prominent feature of genome evolution. The initial focus of their study is on the shared presence and chromosomal distributions of genes. The approach is based on the identification of shared orthologous genes that are used to compare and contrast genome architectures. Varied approaches based on this idea have been used for many comparisons. For instance, human and chimpanzee last shared a common ancestor ~4 million years ago (mya) (Hobolth et al. 2007). Their proximity to one another is reflected in highly similar genome architectures, with changes indicative of species-specific differences (Chimpanzee Sequencing Consortium 2005). At 75 My of divergence, the human and mouse genomes are closely related with ~80% orthologous gene content (Waterston et al. 2002). Since they diverged, rearrangements have reshaped the genomes into 281 syntenic regions of at least 1 Mb (Pevzner and Tesler 2003). Conservation has also been detected in unicellular eukaryotes. The parasitic kinetoplastid genera *Leishmania* and *Trypanosoma* last shared a common ancestor between 200 and 500 mya, predating the divergence of mammals. Despite considerable time, it has been shown that *L. major* and *T. brucei* have maintained ~70% of their genomes in conserved syntenic blocks due to their use of polycistronc transcription (El-Sayed et al. 2005). Synteny can also be detected over surprising evolutionary distances. The chordate amphioxus is a marine animal that shares many developmental and anatomical features with vertebrates. Despite more than half a billion years separating humans and amphioxus, 1,044 genes have been maintained in conserved microsyntenic blocks (Putnam et al. 2008).

Apicomplexan Genome Architectures

Estimates of the last apicomplexan common ancestor range from ~350–824 mya (Escalante and Ayala 1995), with more recent estimates narrowing the range to approximately ~420 mya Berney and Pawlowski 2006; Okamoto and McFadden 2008). An historic lack of whole-genome sequence data has left comparative genomics investigations within the phylum relatively limited, with foci on single genera. Genome structure comparisons within the most infamous apicomplexan genus, *Plasmodium* (the causative agent of malaria), have revealed a high degree of synteny within the genus. This is despite the fact that only ~85% of the genes display orthology (Kooij et al. 2005). Within the genus *Theileria* (a cattle parasite of veterinary importance), the level of synteny is extremely high, with differences primarily due to species-specific genes and the unequal expansion of multicopy gene families (Pain et al. 2005). In contrast, a cross-genera comparison of the genera *Plasmodium* and *Theileria* has revealed rearrangements so extensive that synteny has been nearly obliterated since they diverged ~100 mya (Brayton et al. 2007).

The recent completion of several apicomplexan genome sequences permits the first detailed investigation of genome evolution within the phylum. We compare 12 species: five species of *Plasmodium*, two species of *Theileria*, *Babesia bovis*, two species of *Cryptosporidium*, and the closely related Coccidia, *T. gondii*, and *Neospora caninum* (fig. 1). Previous studies within the phylum indicated that there were sufficient protein-encoding orthologs available to undertake this investigation (Kuo and Kissinger 2008). The 12 species that are available are not uniformly distributed throughout the phylum and represent only four evolutionary lineages, the Haemosporidia (*Plasmodium*), Piroplasmida (*Theileria/Babesia*), Coccidia (*Toxoplasma/Neospora*), and a gregarine-related lineage (*Cryptosporidium*). The analyses are based on current genome assemblies and annotations. The data used are certain to contain physical gaps and misannotated genes. Although the trends and patterns observed here are not likely to change with improved annotation, care must be taken with fine-scale interpretations.

A bioinformatics pipeline to identify orthologous genes, calculate syntenic regions between each pair of genomes, and visualize the results was constructed to determine the changes in genome architecture that have occurred since these species last shared a common ancestor. We show that high levels of syntenic conservation are detected within each of the four lineage groups. Species-specific genes and the expansion of multicopy gene families occur to varying degrees at sites of genome rearrangement. Synteny between the four major lineage groups has been nearly obliterated. These changes have occurred despite the apparent absence of active, rearrangement mediating, TEs in any of the investigated genomes.

Materials and Methods

Data Harvesting, Formatting, and Ortholog Clustering

All data represent the most up to data release at the time of analysis. Annotated protein-encoding genes (along with their genomic coordinates) and the sizes and numbers of chromosomes/scaffolds/contigs for each species were obtained as follows: *P. falciparum*, *P. vivax*, *P. knowlesi*, and *P. chabaudi* data were downloaded from PlasmoDB (Aurrecoechea et al. 2009) version 6.3. *Cryptosporidium muris* and *C. parvum* data were downloaded from CryptoDB (Heiges et al. 2006) version 4.3. The 45 scaffolds from *C. muris* are not assigned to chromosomes and are numbered 1–45. *Toxoplasma gondii* data were obtained from ToxoDB (Gajria et al. 2008) version 6.0. The unpublished *N. caninum* sequence data were obtained from http://www.sanger.ac.uk/resources/downloads/protozoa/neospora-caninum.html. *Babesia bovis* data were downloaded form NCBI (http://www.ncbi.nlm.nih.gov/). Data from chromosomes 1 and 4 were present in multiple parts (7 and 3, respectively) due to gaps in the genome assembly. Accession numbers for these sequences are: Chromosome 1: AAXT01000005, AAXT01000006, AAXT01000008, AAXT01000009, AAXT01000010, AAXT01000011, and AAXT01000012; Chromosome 2: NC_010574; Chromosome 3: NC_010575; and Chromosome 4: AAXT01000002, AAXT01000004, and AAXT01000013. *Plasmodium berghei* data were downloaded from the April 2010 release at Wellcome Trust Sanger Institute (http://www.sanger.ac.uk/Projects/Pathogens/). *Theileria annulata* data were downloaded from the Wellcome Trust Sanger Institute (http://www.sanger.ac.uk/Projects/Pathogens/) on 11th August 2008. *Theileria parva* data were downloaded from TIGR Eukaryotic Genome Projects (ftp://ftp.tigr.org/pub/data/Eukaryotic_Projects/t_parva/annotation_dbs/) on 11th August 2008.

Orthologous gene clusters were identified using a combination of WU-BLAST (http://blast.wustl.edu/) (version 2.2.6, *E* value cutoff of 1×10^{-30}) for an all-by-all BLASTP similarity search (all annotated protein-encoding genes from all species were analyzed) and OrthoMCL (Li et al. 2003) (version 1.4) with default parameters. OrthoMCL uses the similarity information from the all-by-all BLASTP to calculate orthologs based on reciprocal best-hit information and also employs an additional step of Markov Clustering (Van Dongen 2000) to improve sensitivity and specificity. Custom PERL scripts were used to query OrthoMCL output and construct sets of multicopy (at least two paralogs), species-specific, and core conserved genes. Species-specific genes have no orthologs. Core genes have at least one ortholog in all species examined.

Synteny Calculation and Visualization

All orthologs identified by OrthoMCL were subsequently compared with each other via an all-by-all BLASTP (Altschul et al. 1990) to generate the appropriate input for the MCSCAN algorithm (Tang et al. 2008) (BLASTP version 2.2.20, *E* value cutoff of 1×10^{-5}). A python script contained in the MCSCAN package was used to filter the BLASTP output to remove self-matches and to reorder the list of resulting gene pairs lexicographically for input into MCSCAN. MCSCAN (version 0.8) was used to calculate synteny between all combinations of genomes using the pooled BLASTP output and the genomic coordinates. MCSCAN was originally developed for use in plant genomes. Some parameters were altered to reflect the smaller size of the apicomplexan genomes (Tang H, personal communication) A less stringent minimum of three genes was required to constitute a syntenic block (default MCSCAN value is 5), with a 25 kb search window used to look upstream and downstream for the next potential syntenic ortholog. The size of this search window is calculated by MCSCAN based on the average intergenic distance in the genomes being compared. Default values were used for all other parameters. Each syntenic block is assigned an *E* value by MCSCAN. The *E* value is a calculation of the likelihood that a detected syntenic block is due to chance. The program authors suggest a cutoff value of 1×10^{-10}. We used a less stringent 1×10^{-5} in order to guard against false-negative results. Individual *E* values for each syntenic block can be found in supplementary table 1, Supplementary

Material online. An expanded search window of 250 kb was used in a separate analysis to look for additional, physically distant, members of syntenic blocks. For tests with randomized gene orders, coordinate information was maintained, and gene IDs were randomized separately for each organism. A pseudorandom number was assigned to each gene ID. IDs were sorted from smallest to largest, while the chromosome IDs and coordinates remained fixed.

Custom PERL scripts were used to parse MCSCAN output and calculate the total number of syntenic blocks, percent of each proteome observed as markers (i.e., found in syntenic blocks), the locations of syntenic break points (SBPs) and the total number and sizes of gaps between syntenic blocks for each combination of genomes. Contigs not incorporated into chromosome assemblies at the time of this analysis were visualized only if they contained syntenic blocks. MCSCAN output was parsed to create files appropriately formatted for input to Circos (Krzywinski et al. 2009) for visualization. The presence of genes within syntenic gaps was calculated based on gene coordinates, and the coordinates of SBPs calculated by MCSCAN.

Organellar Targeting

Plasmodium falciparum genes were chosen for further study because of the advanced state of the *P. falciparum* annotation relative to the other species and because many of the available prediction tools were developed specifically for use with this organism. Sequences and gene IDs for the 88 *P. falciparum* genes (see "Limited synteny between lineages" in Results) were extracted from PlasmoDB along with available gene product and annotated gene ontology information. This information was searched to identify potential patterns shared by the genes that have remained syntenic. Inspection revealed many genes with products associated with organelles. Search tools at PlasmoDB were used to compare gene IDs with those of 388 genes encoding proteins with subcellular localization evidence placing them in the apicoplast.

The 88 sequences were used as input for three programs designed to predict targeting signals. PATS (version 1.2.1N, from PlasmoDB) (Zuegge et al. 2001) predicts apicoplast targeting. PATS returns two results; a binary "yes" or "no" to indicate the likelihood that a sequence is targeted and a score that ranges from 0 to 1. All sequences with a yes were accepted. Scores for these sequences ranged from 0.546 to 0.997. Plasmit (from PlasmoDB) (Bender et al. 2003) predicts mitochondrial targeting with scores from 1% to 100%. All predicted sequences scored greater than 90%. Predotar (Small et al. 2004) predicts targeting to both the apicoplast and mitochondria with scores ranging from 0 to 1. All sequences with scores above 0.2 were accepted according to the documentation.

Sequences with no other evidence were individually screened with additional tools. PlasmoAP (Foth et al. 2003) predicts apicoplast targeting by predicting both the signal and transit peptides (See Discussion). The program Signal P (Bendtsen et al. 2004) is used to predict the signal peptide. The algorithm's final decision is based on the presence of both peptides, with the highest scoring designation indicating that the sequence is "very likely" to be targeted. None of the remaining genes contain a predicted signal peptide. However, there were some cases where the presence of a transit peptide was predicted with the highest likelihood. These instances (6 total) are noted in supplementary table 2, Supplementary Material online but were not considered as evidence of targeting. MitoProt (Claros and Vincens 1996) predicts targeting to the mitochondria based on a score ranging from 0 to 1 and a cleavage site prediction. All MitoProt predictions (3 total) had a score of at least 0.87 and a predicted cleavage site.

Results

Terminology

Orthologous genes share a common ancestor but are found in different species as a result of speciation. All comparisons in this study are based on the identification of orthologs from the pooled annotated protein-encoding gene sequences of each species. OrthoMCL identifies both orthologs and paralogs (within genome duplications) and outputs both as "ortholog clusters." Only identified orthologs are useful as potential indicators of synteny because their relative positions can be investigated in different genomes, we did not use the paralogs. We use the term "marker" to represent an ortholog that is detected as part of a syntenic region. Individual syntenic regions are referred to as "blocks" of synteny. The nonsyntenic regions that separate blocks are refereed to as gaps. The specific locations where synteny ceases and a gap begins are referred to as synteny break points (SBPs). Not all orthologous genes are found in syntenic regions. Genes with orthologs in all investigated species are called "core" genes. Genes with at least two copies in a single genome are called "multicopy." Genes with no orthologs are called "species-specific."

Detecting Orthologous Genes

All annotated protein sequences from 12 apicomplexan species (fig. 1) were obtained as described in Materials and Methods. The program OrthoMCL (Li et al. 2003) was used to cluster orthologs. Clustering identified a minimum core set of 874 homologous gene clusters, containing at least one gene from all 12 species (10,726 genes in total, including paralogs) (data not shown). Only clusters containing genes from all species are useful for identifying synteny across the entire phylum, however, all orthologous genes were used to detect synteny for individual pairwise species comparisons not only the core set of 874 clusters. Orthologs between lineages primarily consist of core genes. For example, of the 1,043 clusters of genes shared between *C. parvum* and *B. bovis*, 874 of them are part of the core set. If each core cluster contained only one gene per species, the core set would contain 10,488 genes (874 clusters × 12 species). Since 10,726 total genes are found in the core clusters, the genes shared by all Apicomplexa are predominantly single copy, or their paralogs have diverged beyond detection. This core gene set consists of between 887 and 919 genes in each species (*C. muris* and *T. gondii*, respectively) and represents between ~11.5% and ~24.3% of the protein-encoding genes (*T. gondii* and *B. bovis*, respectively)

Table 1. Numbers of Orthologous Gene Clusters Between Species[a].

Taxon	Babesia bovis	Cryptosporidium muris	Cryptosporidium parvum	Neospora caninum	Plasmodium berghei	Plasmodium chabaudi	Plasmodium falciparum	Plasmodium knowlesi	Plasmodium vivax	Theileria annulata	Toxoplasma gondii	Toxoplasma parva
B. bovis	—	1,079	1,043	1,448	1,507	1,500	1,516	1,499	1,502	2,217	1,455	2,208
C. muris		—	2,928	1,434	1,237	1,229	1,240	1,234	1,228	1,069	1,443	1,065
C. parvum			—	1,380	1,189	1,182	1,190	1,187	1,180	1,037	1,387	1,035
N. caninum				—	1,836	1,828	1,852	1,835	1,836	1,427	6,255	1,423
P. berghei					—	4,587	4,307	4,271	4,239	1,499	1,851	1,487
P. chabaudi						—	4,292	4,255	4,226	1,490	1,842	1,478
P. falciparum							—	4,318	4,296	1,506	1,865	1,495
P. knowlesi								—	4,582	1,487	1,851	1,475
P. vivax									—	1,492	1,850	1,482
T. annulata										—	1,434	3,133
T. gondii											—	1,429
T. parva												—

[a] Numbers of orthologous protein-encoding gene clusters used to detect synteny for each species pair.

annotated in each genome. All numbers of ortholog clusters used for pairwise species synteny detection are presented in table 1.

Synteny Detection: Rapidly Drifting Genomic Landscapes

Genomic coordinates for all clustered orthologs (chromosome, scaffold, or contig positions) were input into MCSCAN. MCSCAN uses coordinate data, along with sequence similarity statistics, to simultaneously calculate blocks of syntenic markers shared between all combinations of genomes. A minimum of three genes was required to make a syntenic block, and a search window of 25 kb was used to look up- and downstream of each block for other potential syntenic markers. A "small" search window (25 kb), relative to model eukaryotes, was used because apicomplexan genomes are very compact and introns, when present, are small. An expanded search window and randomized gene orders were used to evaluate these parameters (see below). Within each of the four major lineages, there is a varying, but high degree, of synteny between species. Table 2 is the first quantitative representation of the degree of syntenic conservation across the phylum and the first indication that extensive rearrangement has occurred within the Apicomplexa. Despite the presence of many orthologous genes, syntenic blocks were conspicuously absent across the entire phylum. With the exception of a relatively limited number of syntenic blocks between *Plasmodium* and *Theileria*/*Babesia*, there are no blocks of synteny of at least three genes conserved across all lineages.

Syntenic blocks were allowed to contain intervening nonsyntenic genes. Therefore, not all genes located within the boundaries of a syntenic block are markers. To investigate the number of genes actually conserved in synteny, the percent of each proteome identified as markers was calculated for each comparison (table 3). Where the degree of synteny is high, the majority of the protein-encoding genes are maintained in conserved blocks. However, as seen in table 3, between *Plasmodium* and *Theileria*/*Babesia* only ~2% or less of the genes are found in syntenic blocks, and there is no observed synteny between the other lineages. Given the parameters used, rearrangements between lineages have been sufficient to completely remove blocks of any three genes in the same order and within 25 kb of each other since they last shared a common ancestor.

The program Circos (Krzywinski et al. 2009) was used to visualize the comparisons made in this study (figs. 2–7). Lines crossing the interior of each circle indicate that synteny is detected. The thickness of each line is indicative of the size/span of the syntenic block. When all species are visualized (fig. 2), the lack of synteny (with the limited exceptions discussed above) is easily observed. There are many lines connecting species within each lineage (compare with the numbers of blocks and percent of each proteome conserved in tables 2 and 3, respectively). However, with the exception of a few relatively small spans between *Plasmodium* and *Theileria*/*Babesia* species, there are no lines that cross the middle of the circle connecting all

Table 2. Number of Syntenic Blocks Between Species.

	Plasmodium	*Plasmodium*	*Plasmodium*	*Plasmodium*	*Plasmodium*	*Theileria*	*Theileria*	*Babesia*	*Cryptosporidium*	*Cryptosporidium*	*Toxoplasma*
Taxon	*falciparum*	*vivax*	*knowlesi*	*berghei*	*chabaudi*	*annulata*	*parva*	*bovis*	*parvum*	*muris*	*gondii*
P. falciparum	—	48 (4,253*)[a]	100 (4,269)	50 (4,226)	52 (4,201)	15 (88)	14 (83)	12 (72)	0[b]	0	0
P. vivax		—	84 (4,514)	39 (4,196)	43 (4,176)	13 (77)	11 (71)	14 (85)	0	0	0
P. knowlesi			—	88 (4,217)	92 (4,197)	12 (73)	11 (70)	13 (82)	0	0	0
P. berghei				—	27 (4,590*)	15 (90)	16 (98)	12 (78)	0	0	0
P. chabaudi					—	15 (90)	17 (104)	12 (79)	0	0	0
T. annulata						—	8 (3,102*)	107 (2,053)	0	0	0
T. parva							—	103 (2,012*)	0	0	0
B. bovis								—	0	0	0
C. parvum									—	60 (2,856*)	0
C. muris										—	0
T. gondii											—

[a] Numbers in parentheses are the total number of gene markers observed in synteny in all blocks. Rarely, a marker was included in multiple, slightly overlapping blocks. In these cases, the number of markers for each species was slightly different. This difference was never more than six markers. These cases are marked with an "*," and the number of markers for the top species is shown.
[b] A "0" indicates no detected synteny.

Table 3. Percentage of Proteomes as Markers in Syntenic Blocks[a].

	Plasmodium	*Plasmodium*	*Plasmodium*	*Plasmodium*	*Plasmodium*	*Theileria*	*Theileria*	*Babesia*	*Cryptosporidium*	*Cryptosporidium*	*Toxoplasma*
Taxon	*falciparum*	*vivax*	*knowlesi*	*berghei*	*chabaudi*	*annulata*	*parva*	*bovis*	*parvum*	*muris*	*gondii*
P. falciparum	—	78.30 78.06	83.54 78.39	86.21 77.60	82.40 77.14	2.32 1.62	2.06 1.52	1.96 1.32	0[b]	0	0
P. vivax		—	88.34 83.10	85.60 77.25	81.91 76.88	2.03 1.42	1.76 1.31	2.32 1.56	0	0	0
P. knowlesi			—	86.03 82.52	82.33 82.13	1.93 1.43	1.74 1.37	2.23 1.60	0	0	0
P. berghei				—	90.03 93.60	2.37 1.84	2.44 2.00	2.12 1.59	0	0	0
P. chabaudi					—	2.37 1.77	2.59 2.04	2.15 1.55	0	0	0
T. annulata						—	77.11 81.96	55.92 54.14	0	0	0
T. parva							—	54.81 49.99	0	0	0
B. bovis								—	0	0	0
C. parvum									—	72.60 75.03	0
C. muris										—	0
T. gondii											—

[a] Percentages are based on the total number of protein-encoding genes in each genome. The upper value is the percent for the taxa in the top row and the lower number is the percent for the taxa in the leftmost column.
[b] A "0" indicates no detected synteny.

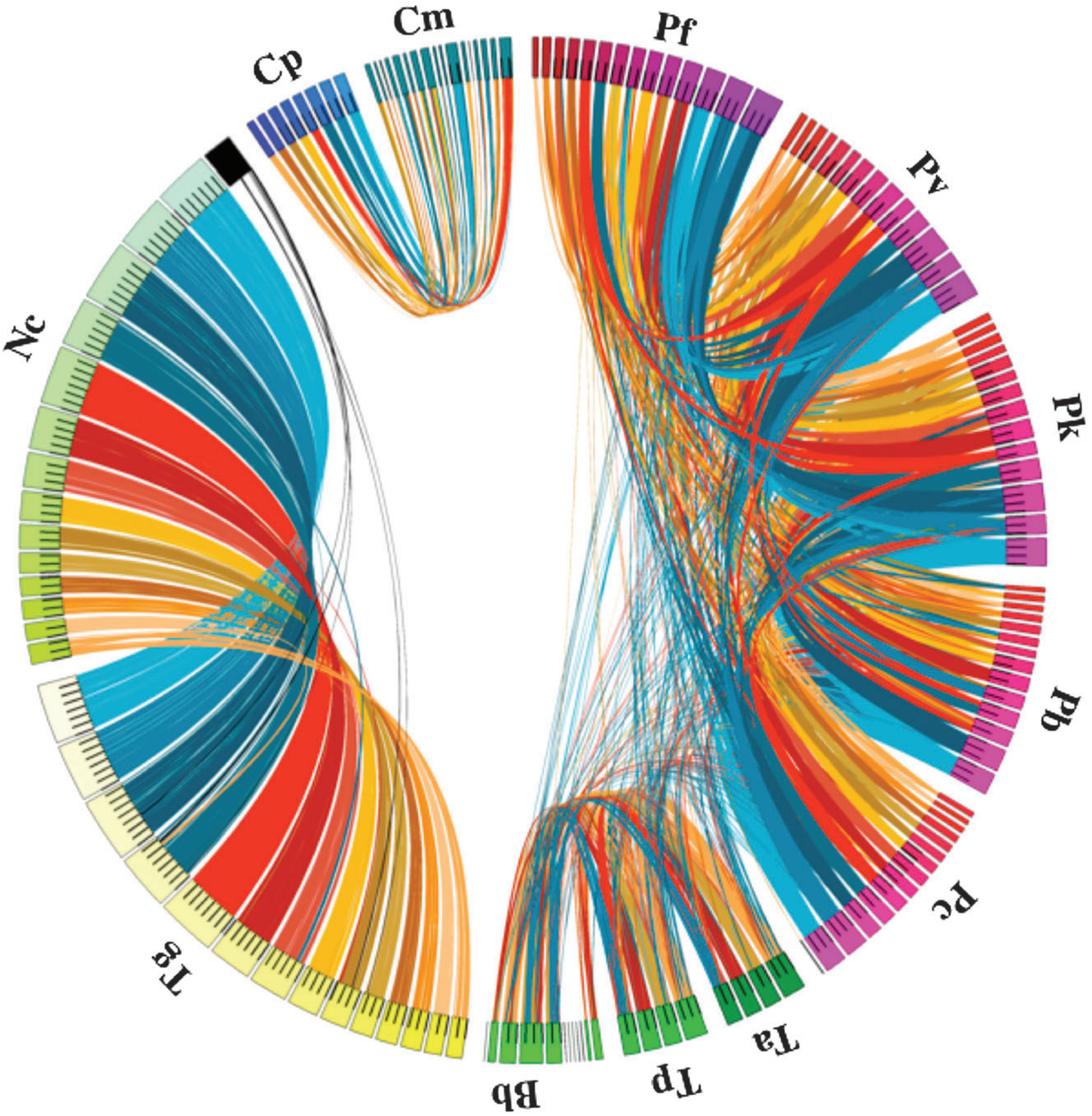

Fig. 2. Detected synteny across the Apicomplexa. The circle is a graphical representation of the annotated chromosomes and contigs in each genome. Each species' genome is labeled with the genus species abbreviation. Scaffolds/Contigs that are not assigned to chromosomes but contain syntenic regions are shown in black. Tick marks represent 1 Mb. Lines that span the interior of the circle connect syntenic regions as detected by MCSCAN. "Twisted" spans represent inversions. Different colors represent different chromosomes within each species.

species. Rapid evolution of genome structure has occurred in the Apicomplexa, leading to extremely different genomic landscapes within the phylum, despite many species having the same number of chromosomes.

Within the genus *Plasmodium* (fig. 3), the conservation of synteny between species recapitulates the phylogeny shown in figure 1. The overall degree of conservation is high, being highest between the most closely related species (as shown by the larger spans with fewer breaks in synteny between the most closely related species). Individual rearrangement events can be tracked by eye. For all comparisons within *Plasmodium*, synteny does not extend to the chromosome ends, which are known to contain species-specific genes involved in host immune evasion (Carlton et al. 2002; Gardner et al. 2002; Kooij et al. 2005; Carlton et al. 2008).

Theileria annulata and *T. parva* are extremely syntenic (fig. 4A). Chromosome 3 has experienced one large and two small intrachromosomal rearrangement events. There is a single interchromosomal event with a small syntenic block (five genes in each species) on *T. annulata* chromosome 1 and *T. parva* chromosome 4. This block contains hypothetical proteins, subtelomeric *Theileria*-specific proteins, and an ATP-binding cassette transporter from each species (supplementary table 1, Supplementary Material

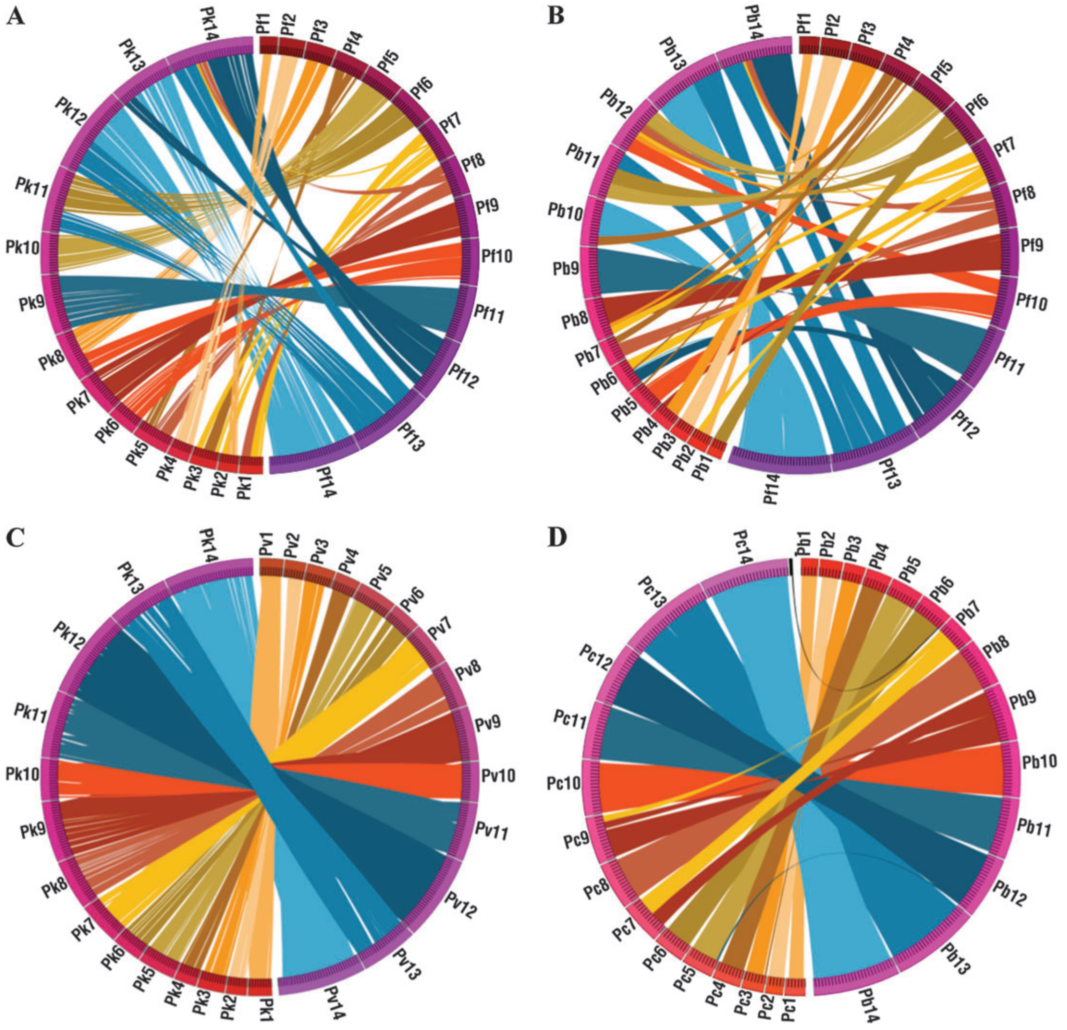

Fig. 3. Synteny between *Plasmodium* species. Each circle represents synteny between two species. Ticks = 100 kb. (*A*) *P. falciparum* and *P. knowlesi*, (*B*) *P. falciparum* and *P. berghei*, (*C*) *P. vivax* and *P. knowlesi*, and (*D*) *P. berghei* and *P. chabaudi*. Chromosome numbers are indicated following the species abbreviation. Different colors represent different chromosomes within each species.

online). Rearrangement between *B. bovis* and both *Theileria* genomes has been more extensive (the *T. parva* relationship to *B. bovis* is virtually identical to that of *T. annulata*) with large chromosomal segments shuffled between genomes.

The most extensive rearrangement within a single genus is found in *Cryptosporidium* (fig. 5). Despite the fact that *C. muris* scaffolds are not assigned chromosome designations, it is apparent that most contain large regions corresponding to multiple *C. parvum* chromosomes. See *C. muris* scaffolds 18, 24, 34, and 43. There are a total of 45 *C. muris* scaffolds. Of the 28 scaffolds (with 62 genes total) where no synteny was found, only four have the minimum three genes required to make a block. The *C. muris* genome is unpublished, and these rearrangements (while not expected to change appreciably) are provisional based on the current assembly and annotation.

Not unexpectedly, extensive synteny is observed between *T. gondii* and *N. caninum* (fig. 2). *Toxoplasma gondii* and *N. caninum* are separated by only ~12 My (Su et al. 2003). A detailed comparison of conserved synteny between these species is in preparation (Reid AJ, Sohal A, Harris D, Quail M, Sanders M, Berriman M, Wastling JM, and Pain A, unpublished data).

Limited Synteny Between Lineages

Further investigation of the few syntenic blocks between *Plasmodium* and *Theileria*/*Babesia* (tables 2 and 3, fig. 6, and supplementary table 1, Supplementary Material online) revealed that the numbers of conserved blocks and genes are similar for all comparisons between lineages. Also, many of the same genes are conserved between species in the two lineages. For example, most *P. falciparum* markers shared with *T. annulata* are also shared with *T. parva* and *B. bovis* (supplementary table 1, Supplementary Material online), and this is unlikely the result of chance. An examination of available gene product information for these genes revealed many genes with a putative role in the mitochondria or apicomplexan plastid, the apicoplast. Within the Apicomplexa, these organellar genomes are streamlined and encode few proteins because the majority of the genes have been lost or transferred to the nuclear genome. Both the apicoplast organelle and mitochondrial genome have been lost in *Cryptosporidium*. Organellar proteins encoded in the nuclear genome are imported into the organelles after translation. There are several well-tested tools designed

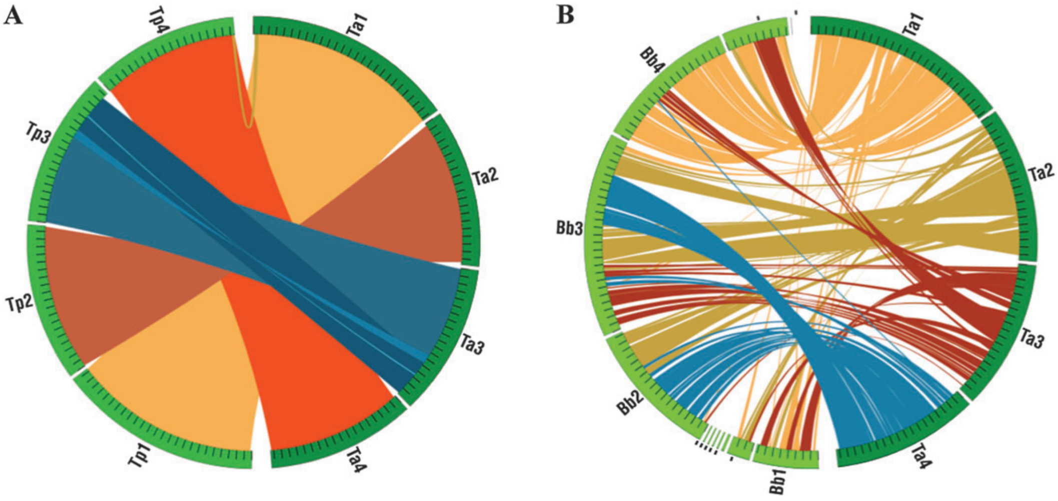

FIG. 4. Synteny between *Theileria* and *Babesia* species. Each circle represents synteny between two species. Ticks = 100 kb. (*A*) *T. annulata* and *T. parva* and (*B*) *T. annulata* and *B. bovis*. Chromosome numbers are indicated following the species abbreviation. Different colors represent different chromosomes within each species. The twisted span connecting chromosome 2 of *T. annulata* and *T. parva* does not indicate an inversion.

to detect targeted genes (see Materials and Methods and Discussion).

The *P. falciparum* genome annotation is generally better than the other *Plasmodium* species. Also, many of the tools available for targeting predictions were developed specifically for use in *P. falciparum*. *Plasmodium falciparum* genes in syntenic regions with *T. annulata* (15 blocks, each with 5–8 genes, supplementary table 1, Supplementary Material online) were examined using a variety of methods, including available annotation information (see Materials and Methods) for evidence of organellar targeting. Of the 88 genes examined, 43 (~49%) had at least one line of evidence indicating that it was targeted to an organelle (supplementary table 2, Supplementary Material online). All blocks contain multiple putatively targeted genes (of the 15 blocks, one contained 2 and the rest contained 3–6). Seven of the genes have evidence of targeting to both organelles (supplementary table 2, Supplementary Material online).

Lost and Found: Syntenic Break Points

Gaps between syntenic blocks for each pair of genomes were calculated based on the locations of SBPs. The numbers and average sizes of the gaps for all comparisons are shown in supplementary table 3, Supplementary Material online. To investigate features common to gaps, data sets of core conserved orthologs, species-specific, and multicopy genes were generated (fig. 7). Circos images clearly show general trends in the distribution patterns of these

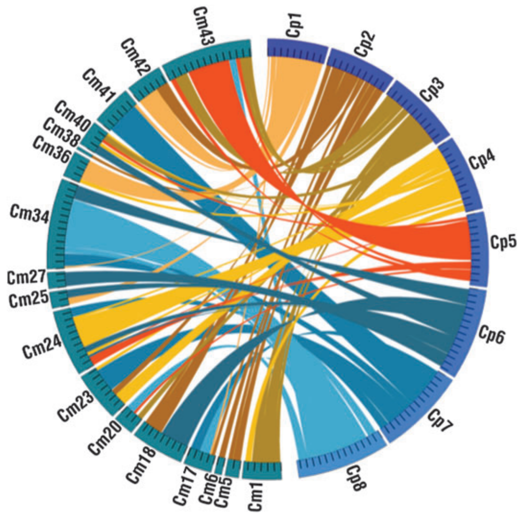

FIG. 5. Synteny between *Cryptosporidium parvum* and *C. muris*. Ticks = 100 kb. *C. muris* contigs are not assigned to chromosomes and are shown in order of designation. Chromosome numbers are indicated following the species abbreviation. Different colors represent different chromosomes within each species.

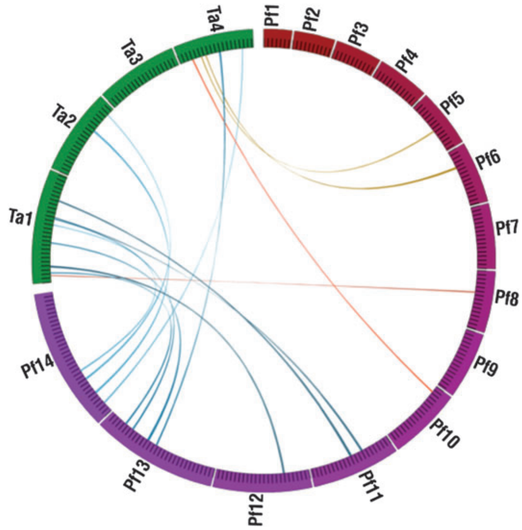

FIG. 6. Limited synteny between *Plasmodium falciparum* and *Theileria annulata*. Ticks = 100 kb. This relationship is representative of the limited synteny between all *Plasmodium* and Piroplasm species. Chromosome numbers are indicated following the species abbreviation. Different colors represent different chromosomes within each species.

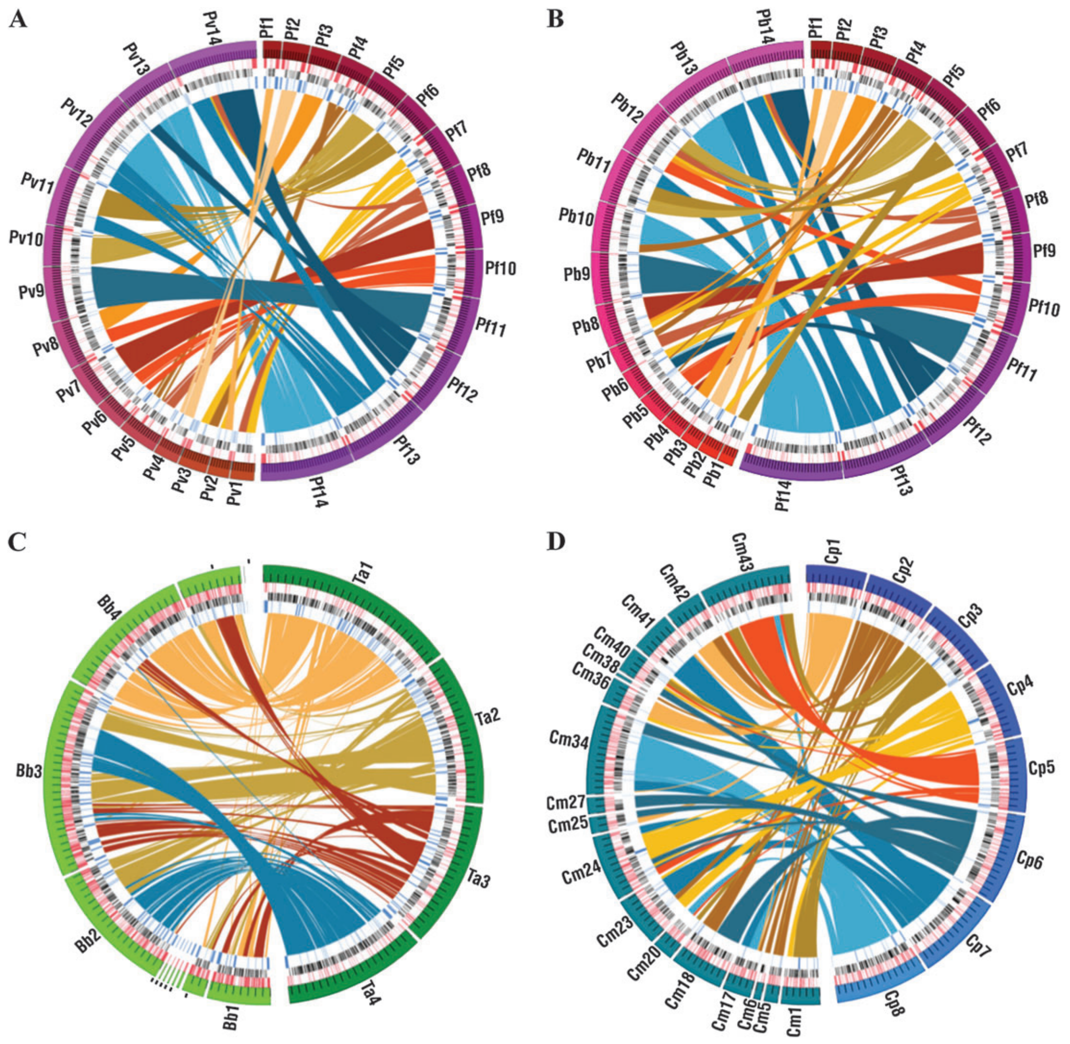

Fig. 7. Distribution of species-specific, core, and multicopy genes. Each circle represents synteny between two species. Ticks = 100 kb. Highlights indicate the position of species-specific (red), core (black) and multicopy (blue) genes for (*A*) *Plasmodium falciparum* and *P. vivax*, (*B*) *P. falciparum* and *P. berghei*, (*C*) *Babesia bovis* and *Theileria annulata*, and (*D*) *Cryptosporidium muris* and *C. parvum*. Chromosome numbers are indicated following the species abbreviation. Different colors represent different chromosomes within each species.

classes of genes. Core genes are mostly absent from chromosome ends (note that *C. muris* contains only scaffolds, therefore, the ends of the molecules in fig. 7 are not necessarily chromosome ends), and otherwise distributed across chromosomes. Species-specific genes are fairly evenly distributed but are concentrated at chromosome ends in *Plasmodium* and to a lesser extent in *Theileria* and *Babesia* (Kuo and Kissinger 2008). Multicopy genes are often species-specific and display a similar distribution pattern (fig. 7). There is a general paucity of species-specific and/or multicopy genes in *P. berghei*, *P. vivax*, and *C. muris*. This is likely a result of incomplete annotations. Species-specific genes can be difficult to annotate, and repetitive gene families can interfere with genome assembly. The abundance of these genes should be considered as a lower limit, likely to increase as annotation and assemblies improve.

Synteny can be disrupted by the generation of novel genes or recombination between members of the same gene family. To explore the possible association of core, species-specific, and multicopy genes with the disruption of synteny, the percent of gaps (including chromosome ends) containing these gene classes was calculated for each pairwise combination of genomes (table 4). Core shared genes are rarely found within gaps and are most common between *Theileria* and *Babesia* (~20% of SBPs). In general, species-specific and multicopy genes are observed more often in gaps than core genes (depending on species, ranges of 11.8–80.6% of gaps contain species-specific genes, and 12–88.9% of gaps contain multicopy genes, table 4). *Theileria parva* and *T. annulata* have nine and eight gaps, respectively (supplementary table 3, Supplementary Material online). The disparity between the numbers of gaps containing species-specific genes for these species (table 4) is probably due to the limited numbers of gaps. Both *Theileria* species have similar relationships with *B. bovis*.

Differences within the genus *Plasmodium* are likely due to a mix of biological factors and annotation effects. *Plasmodium berghei* and *P. chabaudi* (each the other's closest investigated relative, Kedzierski et al. 2002; Perkins et al. 2007) have the lowest number of gaps compared with

Table 4. Percentage of Gaps Containing Core, Species-Specific, and Multicopy Genes[a].

Taxon	*Plasmodium falciparum*	*Plasmodium vivax*	*Plasmodium knowlesi*	*Plasmodium berghei*	*Plasmodium chabaudi*	*Theileria annulata*	*Theileria parva*	*Babesia bovis*	*Cryptosporidium parvum*	*Cryptosporidium muris*	*Toxoplasma gondii*
P. falciparum	—	3.22/3.28 50.0/70.5 41.9/62.3	4.39/4.39 73.7/43.0 74.6/36.0	4.69/3.12 20.3/67.2 42.2/59.4	6.10/4.50 37.9/65.2 36.4/57.6	—	—	—	—	—	—
P. vivax		—	3.06/2.04 80.6/29.6 83.7/23.5	3.77/1.89 26.4/58.5 50.9/47.2	5.26/3.51 45.6/54.4 43.9/43.9	—	—	—	—	—	—
P. knowlesi			—	4.90/4.90 11.8/80.4 27.5/82.4	5.66/5.66 24.5/77.4 23.6/79.2	—	—	—	—	—	—
P. berghei				—	2.43/5.13 65.9/17.9 63.4/61.5	—	—	—	—	—	—
P. chabaudi					—	—	—	—	—	—	—
T. annulata						—	0/0 37.5/77.8 50.0/88.9	22.2/21.9 76.8/45.8 41.4/35.4	—	—	—
T. parva							—	24.0/24.2 77.1/49.5 42.7/38.9	—	—	—
B. bovis								—	—	—	—
C. parvum									—	8.00/12.1 50.7/66.7 12.0/19.7	—
C. muris										—	—
T. gondii											—

[a] Percentages are based on the total number of gaps. For each cell, the top number is core (black), the middle number is species-specific (red), and the bottom number is multicopy (blue). The left value is for the taxa in the top row, and the right value is for the taxa in the left column. Values are not shown for relationships with limited or no synteny.

other *Plasmodium* species, indicating fewer rearrangements (supplementary table 3, Supplementary Material online). *Plasmodium berghei* contains approximately half the number of annotated species-specific and multicopy genes as *P. chabaudi* (data not shown). Although the percentage of gaps containing multicopy genes for each species is nearly identical (table 4), many more gaps contain species-specific genes in *P. chabaudi*. It is possible that there are many unannotated, single copy, species-specific genes in *P. berghei*. Both species show similar synteny patterns relative to other *Plasmodium* species.

From a genome architecture perspective, *P. knowlesi* appears to be evolving more quickly relative to *P. falciparum*, *P. berghei*, *P. chabaudi*, and *P. vivax* than they are to each another. *Plasmodium knowlesi* has approximately half of the annotated species-specific genes and ~100 less multicopy genes compared with its closest investigated relative *P. vivax* (data not shown). However, *P. knowlesi* has approximately twice as many gaps compared with other *Plasmodium* species (supplementary table 3, Supplementary Material online), indicating a rapidly evolving genome architecture. There is a higher percentage of gaps containing both species-specific and multicopy genes when compared with *P. vivax*. This trend extends to all *Plasmodium* species (table 4). *Plasmodium knowlesi* appears to be accumulating gaps and species-specific and multicopy genes within them at a greater rate than other species (as indicated by lower percentages of gaps containing these gene types in all comparisons).

Plasmodium falciparum has the most complete annotation of the *Plasmodium* species sequenced to date. Depending on the comparison (with the exception of the *P. knowlesi* comparison, see above), approximately 60–70% of *P. falciparum* gaps contain species-specific or multicopy genes. There has been an accumulation of these genes since these species last shared a common ancestor, further highlighting their possible role in genome evolution.

Expanded Search Window and Gene Order Randomization

To test the reliability of our search methods, we altered two parameters in our test procedure. MCSCAN searches a set distance up- and downstream of a syntenic region to find the next possible syntenic gene. The size of the search window is based on intergenic distances in the investigated organisms (see Materials and Methods). To guard against false negatives (i.e., the possibility that more synteny is present and was not detected because the search window was too narrow), we increased the search window by an order of magnitude to the MCSCAN default value for plant genomes (see Materials and Methods). To guard against false positives (i.e., the possibility that the observed synteny is due to the chance ordering of genes), the order of the genes was randomized. The randomized gene orders were used with both search windows (see Materials and Methods). To test these parameters, one genome from each lineage (*P. vivax*, *T. annulata*, *C. parvum*, and *T. gondii*) was chosen as a representative for comparison with *P. falciparum* (table 5).

When gene orders are randomized, no synteny is detectable with the 25-kb search window used in the

Table 5. Parameter Validation[a]

	Plasmodium vivax			*Theileria annulata*			*Cryptosporidium parvum*		*Toxoplasma gondii*	
	25 kb	250 kb	R250 kb	25 kb	250 kb	R250 kb	250 kb	R250 kb	250 kb	R250 kb
Block number	48	33	101	15	147	64	77	54	9	5
% Proteome[b]	78.3	78.07	10.44	2.32	20.70	8.91	9.88	7.46	0.58	0.35
Average block size (kb)	~416	~741	~550	~24.5	~448	~537	~473	~484	~430	~542
Total markers	4253	4241	567	88	785	338	376	284	46	28
Average markers per block	88.60	128.52	5.6	5.86	5.34	5.28	4.89	5.26	5.11	5.6
Average *E* value[c]	1.50×10^{-27}	9.09×10^{-7}	1.87×10^{-6}	7.41×10^{-7}	8.54×10^{-7}	2.21×10^{-6}	2.91×10^{-6}	1.85×10^{-6}	2.82×10^{-6}	8.09×10^{-7}
Median *E* value[c]	3.30×10^{-239}	5.35×10^{-242}	3.60×10^{-7}	8.10×10^{-6}	3.10×10^{-8}	1.10×10^{-6}	1.70×10^{-6}	7.86×10^{-7}	2.30×10^{-6}	8.00×10^{-8}

[a] One species from each of the four major lineages was chosen for comparison to *P. falciparum*. For each comparison, four experiments are summarized: the experimental search window size (25 kb, no synteny detected for *C. parvum* and *T. gondii*), randomized gene order with the experimental search window size of 25 kb (no synteny detected for any comparison, not shown in the table), an expanded search window size (250 kb), and randomized gene order with an expanded search window size (R250 kb).

[b] The percentage of the total number of protein-encoding genes found in syntenic blocks.

[c] *E* values are calculated by MCSCAN for each syntenic block. The lower the *E* value, the less likely that a block was detected due to the chance order of genes (see Materials and Methods). *E* values are for all blocks in each comparison.

experimental analysis. This makes it unlikely that any of the observed syntenic regions in the "standard" parameter set used for this study are due to chance (false positives). Because no synteny was detected with the randomized gene order and the 25 kb search window, there are no results for these conditions shown in table 5. Increasing the search window to 250 kb (table 5) greatly increases average syntenic block sizes in comparisons of *P. falciparum* with *P. vivax* and *T. annulata*. In the case of *P. vivax*, the expanded search window finds nearly the same number of syntenic markers as the standard parameter set (table 5). Syntenic regions are collapsed into fewer larger blocks. Average and median *E* values for the blocks (calculated by MCSCAN, see Materials and Methods) were similar, indicating that the blocks are unlikely to occur by chance with either parameter set. However, the average *E* value is lower for the 25-kb search window, indicating that the blocks are less likely to have occurred by chance. For these closely related species, the expanded search window appears to offer no direct advantage. When *P. falciparum* is compared with the more distant *T. annulata*, the expanded 250 kb search window detected additional and larger regions of synteny. However, this increased sensitivity came at a loss of selectivity. The average block size is nearly the same as the size necessary to detect synteny when gene orders are randomized (25 and R250 kb, table 5). These large blocks (~0.5 Mb) contain only ~5 syntenic genes each. When *P. falciparum* was compared with *C. parvum* or *T. gondii*, no synteny was detected with the standard parameter set, and all statistics are similar with both the expanded and random parameters. Thus, the expanded search window is no more effective than the randomized test for either comparison (table 5). Overall, the expanded search window size of 250 kb offers no advantage for synteny detection implying that no significant synteny is likely to have been missed using our search criteria. Furthermore, synteny is not detected at all when gene orders are randomized and the experimental 25 kb search window is used. Therefore, it is unlikely that the results observed in this study are due to the chance ordering of genes.

Discussion

Dynamic Apicomplexan Genomes

Care must be taken when comparing trends and patterns of genome evolution between different groups of eukaryotes. Such comparisons can traverse vast quantities of evolutionary time, equaled by the range of eukaryotic diversity and lifestyle adaptations. Even in the chordates, where synteny can be detected over long distances (see Introduction), recent findings have shown that there are cases where genome architecture appears to be diverging at an accelerated rate. The tunicate *Oikopleura dioica*, exhibits extensive genome rearrangements relative to ancestral chordate linkage groups, including those shared between amphioxus and human, despite having retained general chordate morphology (Denoeud et al. 2010). The long-term conservation of synteny generally observed in chordates is also observed in some protists, for example, the kinetoplastids. In fact, one major difference between "model" chordates and parasitic protists is that rates of recombination and rearrangement are reported to be generally lower in protists (Kooij et al. 2005), with considerable conservation present over great evolutionary distances in some lineages (El-Sayed et al. 2005; Kooij et al. 2005; Pain et al. 2005; Peacock et al. 2007; Weir et al. 2009). In contrast, the Apicomplexa display an unprecedented degree of genome rearrangement with the near complete removal of synteny between major lineages within the phylum.

There are several possible explanations. The loss of apicomplexan synteny could be partially due to increased recombination and rearrangement rates resulting from short generation times relative to model multicellular eukaryotes. To test this hypothesis, we examined the extensive investigations of synteny that have been carried out in the fungi. The divergence time for the basal Ascomycota–Basidiomycota split is estimated at ~450–1,500 mya (Taylor and Berbee 2006). Within the hemiascomycete yeast lineage (divergence time comparable to chordates at ~300–400 My), syntenic conservation is variable across lineages but extensive and detectable (Fischer et al. 2006; Sherman et al. 2009). Generation times vary considerably within the fungi but can be as low as ~1.5 h for the model organism *Saccharomyces cerevisiae*. Three species of the ascomycete genus *Aspergillus* (divergence time ~200 My) have maintained ~77% of their genomes in synteny (Galagan et al. 2005) despite a cell cycle time of only ~90–120 min for the model organism *A. nidulans* (Bergen and Morris 1983). Within the basidiomycota, the *Coprinopsis cinerea* genome has maintained ~40% of its genome syntenic with *Laccaria bicolor* (last common ancestor ~100–200 mya, *C. cinerea* generation time ~2 weeks) (Stajich et al. 2010). Although generation time is a likely factor in rearrangement rates, synteny is still detectable between eukaryotes with distant relationships and shortened generation times.

Another possibility is that the estimates of divergence times among the Apicomplexa and of the Apicomplexa with respect to other Alveolates are incorrect. Recent estimates place the last common apicomplexan ancestor at ~420 mya (Berney and Pawlowski 2006; Okamoto and McFadden 2008). This timescale is less than the estimated time separating organisms where synteny has been detected. However, it is possible that there are other forces at work that have skewed the current estimates. We may be investigating much older relationships. In this case, it is possible that the degradation of synteny is proceeding according to what can be expected based on previously studied eukaryotes.

Give and Take: Removal and Generation of DNA

Despite the expansion of multicopy gene families and the generation of novel species-specific genes (Pain et al. 2005, 2008; Kuo and Kissinger 2008; Weir et al. 2009), the Apicomplexa have extremely small eukaryotic genomes characterized by gene loss. Overall, the removal of genetic material has outpaced the generation of novel DNA, resulting in small genome sizes. Given the otherwise near universal presence of TEs in all other lineages studied to date and their presence in the closest relatives of the Apicomplexa (the ciliates and dinoflagellates), the most parsimonious explanation is that TEs have been lost from the phylum. It is attractive (though only a possibility) to think that the ability of TEs to promote rearrangements and increase genome size was selected against in a "host" nuclear genome with an already accelerated rate of rearrangement and the apparent evolutionary pressure to keep genome sizes small. The need for innovation (via novel gene formation or the maintenance of lineage- or species-specific genes) may have led to intensive genome scrambling, with species-specific and multicopy genes enriched in gaps. The Apicomplexa may be representative of what genomes "look" like when under pressure to develop and maintain a parasitic lifestyle, innovate in the absence of TEs, and maintain reduced genome sizes.

The apparent selection of genome compaction and streamlining (characteristic of parasite genomes) has been observed across the phylum and is the focus of study in *C. parvum* (Abrahamsen et al. 2004; Keeling 2004). Among the Apicomplexa, *C. parvum* has a particularly compact genome and metabolic repertoire, exemplified by its inability to synthesize nucleotides (Striepen et al. 2004). In the Apicomplexa, genome compaction is partially counterbalanced by *de novo* gene creation. Gene creation has been vital in the development of virulence in the Apicomplexa. Species-specific antigen variation genes are present in multiple copies, especially in *Plasmodium* and Piroplasm species (al-Khedery et al. 1999; Carlton et al. 2002; Gardner et al. 2002, 2005; Pain et al. 2005).

To briefly investigate the incidence of gene creation versus loss, we gathered all ortholog clusters at six levels of the cladogram in figure 1 (following the methods in Kuo and Kissinger 2008). Clusters at each level contain genes with orthologs in all species that extend from that level to the tips of the cladogram. For example, ortholog clusters containing core-conserved genes shared by all Apicomplexa are found at level 1 in figure 1. Clusters specific only to the genus *Plasmodium* are found at level 4. All six levels were investigated for the path leading to *P. falciparum*. At each level, a single *P. falciparum* ortholog was compared with all available nonapicomplexan protein sequences in the Genbank (Benson et al. 2003) (BLASTP version 2.2.22+, *E* value cutoff of 1×10^{-3}, minimum 30% identity over at least 50 amino acids). If a potential ortholog outside the Apicomplexa was discovered, we infer that the gene was lost in the other apicomplexan lineages as opposed to being created. Likewise, species-specific genes with no similarity outside the Apicomplexa are likely to have been generated *de novo* within that lineage or species. The largest percentage of orthologs with nonapicomplexan hits, 96.5%, is observed at level 1, the level shared by all Apicomplexa (supplementary table 4, Supplementary Material online). Thus, most genes shared by all apicomplexan species in figure 1 have orthologs in nonapicomplexan species. At each increasing level (fig. 1 and supplementary table 4, Supplementary Material online), fewer of the genes have nonapicomplexan hits, indicating that more of them were created at those levels. At level 6, the most specific level, only 22.8% have nonapicomplexan hits, indicating that most of these genes were likely generated *de novo*. At the levels examined, there is a clear pattern of selective gene loss closer to the base of figure 1, and an increase in *de novo* gene creation moving toward the tips.

Evidence for the involvement of species-specific and multicopy genes in the evolution of apicomplexan genome architecture continues to grow. Improved annotation and detection methods have revealed a greater enrichment for these categories of genes in gaps than previously detected in *Plasmodium* (Kooij et al. 2005). In addition, many comparisons show a greater percentage of gaps containing these genes (table 4) than has been found in another group of unicellular and parasitic protists, the kinetoplastid trypanosomatids, where only ~40% of gaps are associated with multicopy gene families and TEs (El-Sayed et al. 2005).

There also appear to be differences in the degree of rearrangement within at least one of the major lineages. *Plasmodium knowlesi* has several unique features relative to other the investigated *Plasmodium* species. It contains intrachromosomal telomeric repeats, antigen variation genes distributed over entire chromosomes, and phenotypic and lifecycle differences (Pain et al. 2008). It also has the most rapidly changing genome architecture, accumulating the greatest number of species-specific and multicopy genes in gaps. This excess accumulation was detected despite a less complete genome annotation compared with *P. falciparum*. Taken together, these observations point to a relationship between rearrangements and the creation of new genes. Alternatively, rearrangement of existing genes may contribute to altered regulation due to novel chromosomal positioning and subsequent histone regulation effects and/or the functional localization of genes in regions that are associated three dimensionally within the nucleus (Chaal et al. 2010; van Steensel and Dekker 2010; Sullivan et al. 2006; Gissot et al. 2007; Gissot and Kim 2008; Gondor and Ohlsson 2009; Westenberger et al. 2009; Ponts et al. 2010).

Synteny Between Lineages: Conserved or Caught in the Act

Initially we hypothesized that the *Plasmodium* and Piroplasm lineages were in the process of losing the remnants of their syntenic conservation. However, an examination of available annotation information for the few genes that remain in syntenic blocks led us to suspect that organellar targeting is playing a role in the maintenance of synteny. Organellar targeting to the mitochondria is accomplished by multiple pathways, most commonly an N-terminal signal peptide (Emanuelsson et al. 2001). Apicoplast targeting is similar but relies on a set of adjacent N-terminal bipartite signals, the signal and transit peptides (Waller et al. 2000). Sequence variation in the targeting signals makes alignment-based detection difficult. The approaches that we used to investigate potential targeting (see Materials and Methods) rely instead on the biochemical properties of the signals. In *P. falciparum*, these tools have predicted the targeting of 545 (Foth et al. 2003; Ralph et al. 2004) and 381 (Bender et al. 2003) genes to the apicoplast and mitochondria, respectively. Based on these predictions, ~17% of nuclear-encoded proteins in *P. falciparum* are transported to these organelles.

Within the examined syntenic blocks, we detected a possible enrichment of genes putatively targeted to the apicoplast and mitochondria, relative to what is seen in the overall *P. falciparum* genome. More experimental data on the mechanisms and numbers of genes targeted to organelles will be necessary to determine if there is a statistically significant enrichment in these blocks. If such enrichment exists, there is currently no explanation for why genes targeted to organelles may be spatially conserved in the genome and this situation warrants further investigation. It is possible that gene regulation or coexpression may play a role. However, an examination of expression profiles of the genes in these groups did not reveal any overt commonalities in expression (based on a search of available expression profiles at PlasmoDB). Most of these putatively targeted organellar genes are not included in the ~17% of *P. falciparum* genes known to be targeted to the apicoplast or mitochondria (see above). These genes may have escaped detection because of differences in targeting mechanisms. For example, the 545 genes known to be targeted to the apicoplast (see above) are likely targeted to the stroma. The extent of gene targeting to organellar membranes is unknown (Lim et al. 2009; Agrawal and Striepen 2010) and will likely include additional genes. Some genes showed evidence of targeting to both organelles. Bimodal targeting remains largely unexplored in *P. falciparum*, though there is evidence that it occurs in *T. gondii* (Pino et al. 2007). Ultimately, verification of targeting must rely on more than in silico methods.

Future Directions: Rearranging Expectations

Genome-wide expression data can reveal spatial expression trends. It will be interesting to see if syntenic regions share any such trends. Currently there is no evidence to support this, with the exception of another well-studied group of pathogenic protists, the Kinetoplastida (see Introduction). Kinetoplastids also have short generation times relative to many model eukaryotes. However, unlike the Apicomplexa, they display a high degree of syntenic conservation. In trypanosomatids, ~43% of the gaps between species in separate genera were associated with the termini of directional gene clusters (DGCs) (El-Sayed et al. 2005). DGCs are variably sized tracts of genes that are transcribed as a unit. They are characteristic of and unique to kinetoplastid genomes. This association points toward a strong conservation mechanism for the maintenance of synteny in these genomes (Smith et al. 2007).

Comparisons with the kinetoplastids cannot serve as a reliable basis for comparative investigation of the causes of apicomplexan rearrangements. Although both groups contain pathogenic protists, their similarities end there. In fact, such a comparison serves best to highlight how little is actually known about the number and variety of selective pressures that contribute to genome evolution. As more diverse organisms and factors are pursued, our understanding is continually forced to change and expand to include new phenomena. For example, TEs were long considered to be "junk DNA" that existed within a host genome as purely selfish denizens (Doolittle and Sapienza 1980; Orgel and Crick 1980). With continued research, it has become clear that TEs exemplify more than simply "selfish DNA" in terms of their effects on the host genome (see Introduction). The extensive rearrangement in the Apicomplexa opens a new chapter in the study of TEs and eukaryotic genome evolution. Initially we hypothesized that the absence of TEs would lead to enhanced chromosome stability and limited rearrangements. However, even in the absence of their disruptive influence, we observe more change than expected. How were TEs removed from these genomes? What is causing this degree of change in their absence? Genomic repeats are not limited to gene families and TEs. Other types of repetitive DNA can also play significant roles in the evolution of genome structure. A systematic and comprehensive investigation of the "repeatomes" of the Apicomplexa will be necessary to fully explore their role in structural genome evolution.

Conclusions

There are different criteria governing genome evolution within the Apicomplexa relative to other well-studied unicellular and multicellular eukaryotes. As additional data are gathered from diverse species, we will be forced to reexamine our assumptions and beliefs about how genomes evolve. Our findings do not apply to all protists, all parasites, or even all organisms with short reproduction times, suggesting that different evolutionary mechanisms and forces predominate in genome evolution in different areas of the tree of life.

Supplementary Material

Supplementary tables S1–S4 are available at *Molecular Biology and Evolution* online (http://www.mbe.oxfordjournals.org/).

Acknowledgments

We thank Ganesh Srinivasamoorthy and Chih-Horng Kuo for assistance with scripting, Lilach Sheiner and Boris Striepen for the detection of organellar targeting, and Taoran Dong and Josh Bridgers for pipeline development and Haibao Tang for MCSCAN. Genome sequence data for *P. berghei*, *P. chabaudi*, and *N. caninum* were produced by the Parasite genomics group at the Wellcome Trust Sanger Institute, provided by Drs Jonathan Wastling and Matt Berriman and are available from http://www.sanger.ac.uk/resources/downloads/protozoa/. We thank The National Institutes of Health for funding this work (award number R01AI068908 to J.C.K.). We also thank the reviewers for improving manuscript clarity.

References

Abrahamsen MS, Templeton TJ, Enomoto S, et al. (20 co-authors). 2004. Complete genome sequence of the apicomplexan, *Cryptosporidium parvum*. *Science* 304:441–445.

Agrawal S, Striepen B. 2010. More membranes, more proteins: complex protein import mechanisms into secondary plastids. *Protist*. 161:672–687.

al-Khedery B, Barnwell JW, Galinski MR. 1999. Antigenic variation in malaria: a 3' genomic alteration associated with the expression of a *P. knowlesi* variant antigen. *Mol Cell*. 3:131–141.

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990. Basic local alignment search tool. *J Mol Biol*. 215:403–410.

Aurrecoechea C, Brestelli J, Brunk BP, et al. (25 co-authors). 2009. PlasmoDB: a functional genomic database for malaria parasites. *Nucleic Acids Res*. 37:D539–D543.

Baldauf SL. 2003. The deep roots of eukaryotes. *Science* 300:1703–1706.

Bartolome C, Maside X, Charlesworth B. 2002. On the abundance and distribution of transposable elements in the genome of *Drosophila melanogaster*. *Mol Biol Evol*. 19:926–937.

Bender A, van Dooren GG, Ralph SA, McFadden GI, Schneider G. 2003. Properties and prediction of mitochondrial transit peptides from *Plasmodium falciparum*. *Mol Biochem Parasitol*. 132:59–66.

Bendtsen JD, Nielsen H, von Heijne G, Brunak S. 2004. Improved prediction of signal peptides: SignalP 3.0. *J Mol Biol*. 340:783–795.

Bennetzen JL. 2000. Transposable element contributions to plant gene and genome evolution. *Plant Mol Biol*. 42:251–269.

Benson DA, Karsch-Mizrachi I, Lipman DJ, Ostell J, Wheeler DL. 2003. GenBank. *Nucleic Acids Res*. 31:23–27.

Bergen LG, Morris NR. 1983. Kinetics of the nuclear division cycle of *Aspergillus nidulans*. *J Bacteriol* 156:155–160.

Berney C, Pawlowski J. 2006. A molecular time-scale for eukaryote evolution recalibrated with the continuous microfossil record. *Proc R Soc B Biol Sci*. 273:1867–1872.

Bohne A, Brunet F, Galiana-Arnoux D, Schultheis C, Volff JN. 2008. Transposable elements as drivers of genomic and biological diversity in vertebrates. *Chromosome Res*. 16:203–215.

Brayton KA, Lau AO, Herndon DR, et al. (28 co-authors). 2007. Genome sequence of Babesia bovis and comparative analysis of apicomplexan hemoprotozoa. *PLoS Pathog*. 3:1401–1413.

Carlton JM, Adams JH, Silva JC, et al. (40 co-authors). 2008. Comparative genomics of the neglected human malaria parasite *Plasmodium vivax*. *Nature* 455:757–763.

Carlton JM, Angiuoli SV, Suh BB, et al. (44 co-authors). 2002. Genome sequence and comparative analysis of the model rodent malaria parasite *Plasmodium yoelii yoelii*. *Nature* 419:512–519.

Chaal BK, Gupta AP, Wastuwidyaningtyas BD, Luah YH, Bozdech Z. 2010. Histone deacetylases play a major role in the transcriptional regulation of the *Plasmodium falciparum* life cycle. *PLoS Pathog*. 6:e1000737.

Chimpanzee Sequencing Consortium. 2005. Initial sequence of the chimpanzee genome and comparison with the human genome. *Nature* 437:69–87.

Claros MG, Vincens P. 1996. Computational method to predict mitochondrially imported proteins and their targeting sequences. *Eur J Biochem*. 241:779–786.

Denoeud F, Henriet S, Mungpakdee S, et al. (56 co-authors). 2010. Plasticity of animal genome architecture unmasked by rapid evolution of a pelagic tunicate. *Science* 330:1381–1385.

Doolittle WF, Sapienza C. 1980. Selfish genes, the phenotype paradigm and genome evolution. *Nature* 284:601–603.

Durand P, Oelofse A, Coetzer T. 2006. An analysis of mobile genetic elements in three *Plasmodium* species and their potential impact on the nucleotide composition of the *P. falciparum* genome. *BMC Genomics* 7:282.

El-Sayed NM, Myler PJ, Blandin G, et al. (44 co-authors). 2005. Comparative genomics of trypanosomatid parasitic protozoa. *Science* 309:404–409.

Emanuelsson O, von Heijne G, Schneider G. 2001. Analysis and prediction of mitochondrial targeting peptides. *Methods Cell Biol*. 65:175–187.

Escalante AA, Ayala FJ. 1995. Evolutionary origin of Plasmodium and other Apicomplexa based on rRNA genes. *Proc Natl Acad Sci U S A*. 92:5793–5797.

Fischer G, Rocha EP, Brunet F, Vergassola M, Dujon B. 2006. Highly variable rates of genome rearrangements between hemiascomycetous yeast lineages. *PLoS Genet*. 2:e32.

Foth BJ, Ralph SA, Tonkin CJ, Struck NS, Fraunholz M, Roos DS, Cowman AF, McFadden GI. 2003. Dissecting apicoplast targeting in the malaria parasite *Plasmodium falciparum*. *Science* 299: 705–708.

Gajria B, Bahl A, Brestelli J, et al. (15 co-authors). 2008. ToxoDB: an integrated *Toxoplasma gondii* database resource. *Nucleic Acids Res*. 36:D553–D556.

Galagan JE, Calvo SE, Cuomo C, et al. (50 co-authors). 2005. Sequencing of *Aspergillus nidulans* and comparative analysis with *A. fumigatus* and *A. oryzae*. *Nature* 438:1105–1115.

Gardner MJ, Bishop R, Shah T, et al. (44 co authors). 2005. Genome sequence of *Theileria parva*, a bovine pathogen that transforms lymphocytes. *Science* 309:134–137.

Gardner MJ, Hall N, Fung E, et al. (45 co-authors). 2002. Genome sequence of the human malaria parasite *Plasmodium falciparum*. *Nature* 419:498–511.

Gissot M, Kelly KA, Ajioka JW, Greally JM, Kim K. 2007. Epigenomic modifications predict active promoters and gene structure in *Toxoplasma gondii*. *PLoS Pathog*. 3:e77.

Gissot M, Kim K. 2008. How epigenomics contributes to the understanding of gene regulation in *Toxoplasma gondii*. *J Eukaryot Microbiol*. 55:476–480.

Gondor A, Ohlsson R. 2009. Chromosome crosstalk in three dimensions. *Nature* 461:212–217.

Heiges M, Wang HM, Robinson E, et al. (13 co-authors). 2006. CryptoDB: a Cryptosporidium bioinformatics resource update. *Nucleic Acids Res*. 34:D419–D422.

Hobolth A, Christensen OF, Mailund T, Schierup MH. 2007. Genomic relationships and speciation times of human, chimpanzee, and gorilla inferred from a coalescent hidden Markov model. *PLoS Genet*. 3:e7.

Hua-Van A, Le Rouzic A, Maisonhaute C, Capy P. 2005. Abundance, distribution and dynamics of retrotransposable elements and transposons: similarities and differences. *Cytogenet Genome Res*. 110:426–440.

Huang J, Kissinger J. 2006. Horizontal and intracellular gene transfer in the Apicomplexa: the scope and functional consequences. In: Katz L, Bhattacharya D, editors. Genome evolution in eukaryotic microbes. New York: Oxford University Press. p. 123–136.

Huang J, Mullapudi N, Lancto CA, Scott M, Abrahamsen MS, Kissinger JC. 2004. Phylogenomic evidence supports past endosymbiosis, intracellular and horizontal gene transfer in *Cryptosporidium parvum*. *Genome Biol*. 5:R88.

Huang J, Mullapudi N, Sicheritz-Ponten T, Kissinger JC. 2004. A first glimpse into the pattern and scale of gene transfer in Apicomplexa. *Int J Parasitol*. 34:265–274.

Kedzierski L, Escalante A, Isea R, Black CG, Barnwell JW, Coppel RL. 2002. Phylogenetic analysis of the genus *Plasmodium* based on the gene encoding adenylosuccinate lyase. *Infect Genet Evol*. 1:297–301.

Keeling PJ. 2004. Reduction and compaction in the genome of the apicomplexan parasite Cryptosporidium parvum. *Dev Cell*. 6:614–616.

Kent WJ, Baertsch R, Hinrichs A, Miller W, Haussler D. 2003. Evolution's cauldron: duplication, deletion, and rearrangement in the mouse and human genomes. *Proc Natl Acad Sci U S A*. 100:11484–11489.

Kooij TW, Carlton JM, Bidwell SL, Hall N, Ramesar J, Janse CJ, Waters AP. 2005. A *Plasmodium* whole-genome synteny map: indels and synteny breakpoints as foci for species-specific genes. *PLoS Pathog*. 1:e44.

Krzywinski M, Schein J, Birol I, Connors J, Gascoyne R, Horsman D, Jones SJ, Marra MA. 2009. Circos: an information aesthetic for comparative genomics. *Genome Res*. 19:1639–1645.

Kuo CH, Kissinger JC. 2008. Consistent and contrasting properties of lineage-specific genes in the apicomplexan parasites *Plasmodium* and *Theileria*. *BMC Evol Biol*. 8:108.

Li L, Stoeckert CJ Jr., Roos DS. 2003. OrthoMCL: identification of ortholog groups for eukaryotic genomes. *Genome Res*. 13:2178–2189.

Lim L, Kalanon M, McFadden GI. 2009. New proteins in the apicoplast membranes: time to rethink apicoplast protein targeting. *Trends Parasitol*. 25:197–200.

Nagamune K, Sibley LD. 2006. Comparative genomic and phylogenetic analyses of calcium ATPases and calcium-regulated proteins in the apicomplexa. *Mol Biol Evol*. 23:1613–1627.

Okamoto N, McFadden GI. 2008. The mother of all parasites. *Future Microbiol*. 3:391–395.

Orgel LE, Crick FH. 1980. Selfish DNA: the ultimate parasite. *Nature* 284:604–607.

Pain A, Bohme U, Berry AE, et al. (54 co-authors). 2008. The genome of the simian and human malaria parasite *Plasmodium knowlesi*. *Nature* 455:799–803.

Pain A, Renauld H, Berriman M, et al. (50 co-authors). 2005. Genome of the host-cell transforming parasite *Theileria annulata* compared with *T. parva*. *Science* 309:131–133.

Peacock CS, Seeger K, Harris D, et al. (42 co-authors). 2007. Comparative genomic analysis of three *Leishmania* species that cause diverse human disease. *Nat Genet*. 39:839–847.

Perkins SL, Sarkar IN, Carter R. 2007. The phylogeny of rodent malaria parasites: simultaneous analysis across three genomes. *Infect Genet Evol*. 7:74–83.

Pevzner P, Tesler G. 2003. Genome rearrangements in mammalian evolution: lessons from human and mouse genomes. *Genome Res*. 13:37–45.

Pino P, Foth BJ, Kwok LY, Sheiner L, Schepers R, Soldati T, Soldati-Favre D. 2007. Dual targeting of antioxidant and metabolic enzymes to the mitochondrion and the apicoplast of. *Toxoplasma gondii*. *PLoS Pathog*. 3:e115.

Ponts N, Harris EY, Prudhomme J, Wick I, Eckhardt-Ludka C, Hicks GR, Hardiman G, Lonardi S, Le Roch KG. 2010. Nucleosome landscape and control of transcription in the human malaria parasite. *Genome Res*. 20:228–238.

Putnam NH, Butts T, Ferrier DE, et al. (38 co-authors). 2008. The amphioxus genome and the evolution of the chordate karyotype. *Nature* 453:1064–1071.

Ralph SA, van Dooren GG, Waller RF, Crawford MJ, Fraunholz MJ, Foth BJ, Tonkin CJ, Roos DS, McFadden GI. 2004. Tropical infectious diseases: metabolic maps and functions of the *Plasmodium falciparum* apicoplast. *Nat Rev Microbiol*. 2:203–216.

Sherman DJ, Martin T, Nikolski M, Cayla C, Souciet JL, Durrens P. 2009. Genolevures: protein families and synteny among complete hemiascomycetous yeast proteomes and genomes. *Nucleic Acids Res*. 37:D550–D554.

Small I, Peeters N, Legeai F, Lurin C. 2004. Predotar: A tool for rapidly screening proteomes for N-terminal targeting sequences. *Proteomics*. 4:1581–1590.

Smith DF, Peacock CS, Cruz AK. 2007. Comparative genomics: from genotype to disease phenotype in the leishmaniases. *Int J Parasitol*. 37:1173–1186.

Stajich JE, Wilke SK, Ahren D, et al. (49 co-authors). 2010. Insights into evolution of multicellular fungi from the assembled chromosomes of the mushroom *Coprinopsis cinerea* (*Coprinus cinereus*). *Proc Natl Acad Sci U S A*. 107:11889–11894.

Striepen B, Pruijssers AJP, Huang J, Li C, Gubbels MJ, Umejiego NN, Hedstrom L, Kissinger JC. 2004. Gene transfer in the evolution of parasite nucleotide biosynthesis. *Proc Natl Acad Sci U S A*. 101:3154–3159.

Su C, Evans D, Cole RH, Kissinger JC, Ajioka JW, Sibley LD. 2003. Recent expansion of *Toxoplasma* through enhanced oral transmission. *Science* 299:414–416.

Sullivan WJ Jr., Naguleswaran A, Angel SO. 2006. Histones and histone modifications in protozoan parasites. *Cell Microbiol*. 8:1850–1861.

Tang H, Wang X, Bowers JE, Ming R, Alam M, Paterson AH. 2008. Unraveling ancient hexaploidy through multiply-aligned angiosperm gene maps. *Genome Res*. 18:1944–1954.

Taylor JW, Berbee ML. 2006. Dating divergences in the Fungal Tree of Life: review and new analyses. *Mycologia* 98:838–849.

Templeton TJ, Enomoto S, Chen WJ, Huang CG, Lancto CA, Abrahamsen MS, Zhu G. 2009. A genome-sequence survey for *Ascogregarina taiwanensis* supports evolutionary affiliation but metabolic diversity between a Gregarine and *Cryptosporidium*. *Mol Biol Evol*. 27:235–248.

Van Dongen S. 2000. Graph clustering by flow simulation. University of Utrecht.

van Steensel B, Dekker J. 2010. Genomics tools for unraveling chromosome architecture. *Nat Biotechnol*. 28:1089–1095.

Waller RF, Reed MB, Cowman AF, McFadden GI. 2000. Protein trafficking to the plastid of *Plasmodium falciparum* is via the secretory pathway. *Embo J*. 19:1794–1802.

Waterston RH, Lindblad-Toh K, Birney E, et al. (222 co-authors). 2002. Initial sequencing and comparative analysis of the mouse genome. *Nature* 420:520–562.

Weir W, Sunter J, Chaussepied M, Skilton R, Tait A, de Villiers EP, Bishop R, Shiels B, Langsley G. 2009. Highly syntenic and yet divergent: a tale of two *Theilerias*. *Infect Genet Evol*. 9:453–461.

Westenberger SJ, Cui L, Dharia N, Winzeler E. 2009. 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*. 10:610.

Zhu G, Keithly JS. 2002. Alpha-proteobacterial relationship of apicomplexan lactate and malate dehydrogenases. *J Eukaryot Microbiol*. 49:255–261.

Zuegge J, Ralph S, Schmuker M, McFadden GI, Schneider G. 2001. Deciphering apicoplast targeting signals–feature extraction from nuclear-encoded precursors of *Plasmodium falciparum* apicoplast proteins. *Gene* 280:19–26.