The Apicomplexan Whole-Genome Phylogeny: An Analysis of Incongruence among Gene Trees

Chih-Horng Kuo,[1] John P. Wares,* and Jessica C. Kissinger*†‡*

*Department of Genetics, University of Georgia; †Center for Tropical and Emerging Global Diseases, University of Georgia; and ‡Institute of Bioinformatics, University of Georgia

The protistan phylum Apicomplexa contains many important pathogens and is the subject of intense genome sequencing efforts. Based upon the genome sequences from seven apicomplexan species and a ciliate outgroup, we identified 268 single-copy genes suitable for phylogenetic inference. Both concatenation and consensus approaches inferred the same species tree topology. This topology is consistent with most prior conceptions of apicomplexan evolution based upon ultrastructural and developmental characters, that is, the piroplasm genera *Theileria* and *Babesia* form the sister group to the *Plasmodium* species, the coccidian genera *Eimeria* and *Toxoplasma* are monophyletic and are the sister group to the *Plasmodium* species and piroplasm genera, and *Cryptosporidium* forms the sister group to the above mentioned with the ciliate *Tetrahymena* as the outgroup. The level of incongruence among gene trees appears to be high at first glance; only 19% of the genes support the species tree, and a total of 48 different gene-tree topologies are observed. Detailed investigations suggest that the low signal-to-noise ratio in many genes may be the main source of incongruence. The probability of being consistent with the species tree increases as a function of the minimum bootstrap support observed at tree nodes for a given gene tree. Moreover, gene sequences that generate high bootstrap support are robust to the changes in alignment parameters or phylogenetic method used. However, caution should be taken in that some genes can infer a "wrong" tree with strong support because of paralogy, model violations, or other causes. The importance of examining multiple, unlinked genes that possess a strong phylogenetic signal cannot be overstated.

Introduction

The protistan phylum Apicomplexa contains many important pathogens (Levine 1988). The most infamous members of this phylum are the causative agents of malaria from the genus *Plasmodium*, which causes more than one million human deaths per year globally (WHO and UNICEF 2005). Other important lineages include *Babesia*, which causes babesiosis in ruminants and humans (Brayton et al. 2007); *Cryptosporidium*, which causes cryptosporidiosis in humans and animals (Abrahamsen et al. 2004); *Theileria*, which causes tropical theileriosis and East Coast fever in cattle (Gardner et al. 2005; Pain et al. 2005); and *Toxoplasma*, which causes toxoplasmosis in immunocompromised patients and congenitally infected fetuses (Montoya and Liesenfeld 2004). These pathogens have been subjected to intense genome sequencing efforts in the hope of facilitating biomedical research (Tarleton and Kissinger 2001; Carlton 2003). The recent availability of fully annotated genome sequences from multiple species within this phylum provides a new and exciting opportunity for us to better understand the phylogeny of these important pathogens.

The use of genome sequences for phylogenetic inference has only recently become possible. The large number of characters derived from genomic data allows robust inference of organismal phylogeny (Delsuc et al. 2005; Philippe, Delsuc, et al. 2005; Rokas 2006), even when the level of incomplete lineage sorting is high (Pollard et al. 2006). Initially, it was thought that use of genomic data would bring an end to the incongruence commonly observed in multigene molecular phylogenetic inference (Gee 2003; Rokas et al. 2003). However, further investigations suggest that the results from genome-scale phylogenetic inference should be interpreted with caution (Soltis et al. 2004; Jeffroy et al. 2006; Nishihara et al. 2007). Although genomic data can effectively suppress stochastic noise in shorter molecular sequences, the large amount of data can actually strengthen systematic biases when present (Phillips et al. 2004; Rodriguez-Ezpeleta et al. 2007).

Previous studies that examined factors such as poor taxon sampling (Soltis et al. 2004; Philippe, Lartillot, and Brinkmann 2005), inappropriate choices of phylogenetic method (Phillips et al. 2004; Jeffroy et al. 2006), nucleotide or amino acid composition bias and deviation from compositional equilibrium (Phillips et al. 2004; Collins et al. 2005), and variation of evolutionary rates among or within sites (Dopazo H and Dopazo J 2005; Nishihara et al. 2007; Rodriguez-Ezpeleta et al. 2007), all found that systematic biases can lead to incorrect trees with strong support. Several approaches that can detect and remove systematic biases in genome-scale phylogenetic inference have been proposed, including modification of taxon sampling (Rodriguez-Ezpeleta et al. 2007), examination of model violations (Rodriguez-Ezpeleta et al. 2007), recoding of molecular sequences (Phillips et al. 2004; Rodriguez-Ezpeleta et al. 2007), removal of the fast-evolving sites (Nishihara et al. 2007; Rodriguez-Ezpeleta et al. 2007), and utilizing rare genomic changes (Delsuc et al. 2005). Among the approaches that have been developed to address the systematic biases in genome-scale analyses, examination of incongruence among individual genes is directly relevant to the design and interpretation of multigene analyses that are fundamental in molecular phylogenetics (Huelsenbeck et al. 1996; Taylor and Piel 2004; Jeffroy et al. 2006). Unfortunately, investigations of incongruence among gene trees at the genome-scale have been limited to a few selected groups such as gamma-Proteobacteria (Lerat et al. 2003), yeast (Taylor and Piel 2004; Gatesy and Baker 2005; Jeffroy et al. 2006), and *Drosophila* (Pollard et al. 2006) due to the limitation of data availability.

[1] Present address: Department of Ecology and Evolutionary Biology, University of Arizona

Key words: Apicomplexa, genome scale, phylogeny, bootstrap, long-branch attraction, taxon sampling.

E-mail: chkuo@email.arizona.edu.

Mol. Biol. Evol. 25(12):2689–2698. 2008
doi:10.1093/molbev/msn213
Advance Access publication September 26, 2008

Table 1
List of Species Name Abbreviations and Data Sources

Abbreviation	Species Name	Data Source[a]	Version Date	Number of Proteins[b]	Genome Size (Mb)
Bb	*Babesia bovis*	GenBank	06 August 2007	3,703	8
Cp	*Cryptosporidium parvum*	CryptoDB.org	13 November 2007	3,805	9
Et	*Eimeria tenella*	GeneDB.org	01 January 2005	11,393	60
Pf	*Plasmodium falciparum*	PlasmoDB.org	24 September 2007	5,460	23
Pv	*Plasmodium vivax*	PlasmoDB.org	24 September 2007	5,352	27
Ta	*Theileria annulata*	GeneDB.org	17 July 2005	3,795	8
Tg	*Toxoplasma gondii*	ToxoDB.org	01 November 2007	7,793	63
Tt[c]	*Tetrahymena thermophila*	J. Craig Venter Institute	04 October 2006	27,424	104

[a] The annotated protein sequences were downloaded from the respective data source with the version date as indicated.
[b] All annotated protein sequences from each species are used to identify single-copy genes that are shared by all species.
[c] The free-living ciliate, *T. thermophila*, is included as the outgroup.

In this study, we present the first genome-scale phylogenetic analysis in the phylum Apicomplexa. Because of the ancient origin of this phylum, estimated at approximately 700–900 Myr (Douzery et al. 2004), we perform our genome-scale phylogenetic inference at the protein level. The robust inference of the organismal phylogeny based on genomic data provides a solid foundation for comparative studies that improve our knowledge of apicomplexan evolution. In addition to facilitating the planning of future phylogenetic studies that involve other closely related pathogens, our systematic investigation of incongruence among gene trees can improve our understanding of multigene phylogenetic inference in general.

Materials and Methods

Data Sources and Ortholog Identification

Our data set contains seven apicomplexan species that have fully annotated genome sequence available, including *Babesia bovis* (Brayton et al. 2007) from GenBank (GenBank accession numbers AAXT01000001–AAXT01000013), *Cryptosporidium parvum* (Abrahamsen et al. 2004) from CryptoDB.org (Heiges et al. 2006), *Eimeria tenella* from GeneDB.org (Hertz-Fowler et al. 2004), *Plasmodium falciparum* (Gardner et al. 2002) and *Plasmodium vivax* from PlasmoDB.org (Bahl et al. 2003), *Theileria annulata* (Pain et al. 2005) from GeneDB.org (Hertz-Fowler et al. 2004), and *Toxoplasma gondii* from Toxo-DB.org (Gajria et al. 2008). A free-living ciliate, *Tetrahymena thermophila* (Eisen et al. 2006), is included as the outgroup. For each species, we obtained all annotated proteins in the genome for ortholog identification. The data sources and protein-encoding gene counts are summarized in table 1.

Orthologous genes were identified using OrthoMCL (Li et al. 2003) (version 1.3) with BLASTP (Altschul et al. 1990) and *E* value cutoff set to 1×10^{-30}. The ortholog identification process in OrthoMCL is largely based on the popular criterion of reciprocal best hits but also involves an additional step of Markov Clustering (van Dongen 2000) to improve sensitivity and specificity. A benchmarking study has found that this algorithm performed well among available methods for ortholog identification (Hulsen et al. 2006). We selected the orthologous genes that are shared by all eight species to infer the gene tree. Orthologous gene clusters that contain more than one gene from any given species were removed to avoid the complications introduced by paralogous genes in phylogenetic inference.

Phylogenetic Inference

The program ClustalW (Thompson et al. 1994) (version 1.83) was used for multiple sequence alignment. The "tossgaps" option was enabled to ignore gaps when constructing the guide tree, and all other parameters were set to the default values unless specifically stated otherwise. The alignments produced by ClustalW were filtered by GBLOCKS (Castresana 2000) (version 0.91b) to using default settings remove regions that contain gaps or are highly divergent. The resulting amino acid alignment for each gene (provided in supplementary data file 1, Supplementary Material online) was used in the main phylogenetic analysis as described below; a codon-based nucleotide alignment for each gene was generated by PAL2NAL (Suyama et al. 2006) and is provided in supplementary data file 2 (Supplementary Material online).

Three phylogenetic methods, including maximum likelihood (ML), maximum parsimony (MP), and Neighbor-Joining (NJ), were used to infer the gene tree for each individual gene. ML inferences were performed using PHYML (Guindon and Gascuel 2003). The proportion of invariant sites and the gamma-distribution parameter with eight substitution categories were estimated from the data set. The substitution model was set to JTT (Jones et al. 1992), and we enabled the optimization options for tree topology, branch lengths, and rate parameters. MP trees were constructed using PROTPARS in the PHYLIP package (Felsenstein 1989) (version 3.65) with 100 randomizations of input order. When more than one equally parsimonious tree was found for a given gene, the strict consensus tree of all equally parsimonious trees was used as the MP tree of this gene. NJ trees were constructed using NEIGHBOR in the PHYLIP package with species input order randomization enabled. The distance matrices were calculated by Tree-Puzzle (Schmidt et al. 2002) (version 5.2). The parameters used in Tree-Puzzle were set to the JTT substitution model, the mixed model of rate heterogeneity with one invariant and eight gamma rate categories, and the exact and slow parameter estimation. The level of bootstrap support for each gene was inferred by 100 resamplings of

the alignment using SEQBOOT in the PHYLIP package followed by ML inference.

To investigate the sensitivity of a gene to the multiple sequence alignment parameter, we varied the gap opening penalty by 2-fold in both directions (i.e., increased the default cost from 10 to 20 or decreased it to 5) and inferred the gene tree under each setting. Individual genes are classified into three categories including robust, intermediate, and sensitive based on the ML gene-tree topologies from the three gap opening penalties examined. A gene is classified as robust if all three settings generated the same topology, intermediate if two out of the three settings generated the same topology, or sensitive if each setting generated a different topology.

To investigate the effect of the substitution model used on the resulting gene-tree topology, we performed ML inference for each gene using two additional substitution models, including LG (Le and Gascuel 2008) and WAG (Whelan and Goldman 2001). The resulting gene trees are compared with the topology obtained using the JTT model (Jones et al. 1992).

Inference of the Species Tree

The species tree was inferred using two different approaches. The first approach was based on the consensus of individual gene trees. The consensus tree was inferred by the CONSENSE program in the PHYLIP package using extended majority rule. Gene trees inferred by different phylogenetic methods (i.e., ML, MP, and NJ) were analyzed separately. The second approach was based on the concatenated alignment of all individual genes following the phylogenetic inference procedures as described above.

Characterization of Gene Trees

The topology distance between each gene tree and the species tree was calculated based on the symmetric difference (Robinson and Foulds 1981) as implemented in TREEDIST in the PHYLIP package. For genes that inferred a topology that is different from the species tree, we performed the approximately unbiased (AU) test (Shimodaira 2002) and the Shimodaira–Hasegawa (SH) test (Shimodaira and Hasegawa 1999) using the CONSEL package (Shimodaira and Hasegawa 2001) to test if the species tree topology is significantly rejected by a gene.

Taxon Removal Tests

To evaluate the potential influence of long-branch attraction (LBA), we removed either of the two taxa that have a long terminal branch (i.e., the outgroup *T. thermophila* and the ingroup *C. parvum*) and repeated the phylogenetic inference for each gene. Our procedure is conceptually similar to the taxon jackknife method (Siddall 1995) but contains one important distinction. The traditional taxon jackknife method removes a taxon after multiple sequence alignment and prior to tree reconstruction. However, the taxon being removed still affects the alignment and thus can influence the resulting tree. We chose to perform the taxon removal prior to multiple sequence alignment to eliminate any effect on the phylogenetic inference from the taxon being removed.

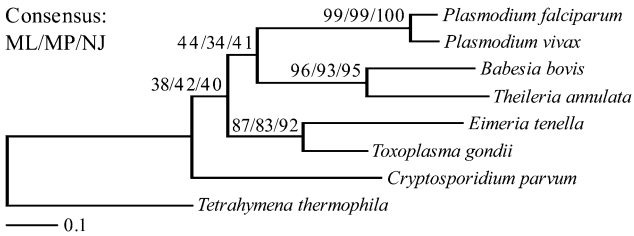

FIG. 1.—The inferred apicomplexan species tree. The ML tree is generated from the concatenated alignment of 268 single-copy genes (71,830 aligned amino acid sites). One free-living ciliate, *Tetrahymena thermophila*, is included as the outgroup to root the tree. Bootstrap support based on 100 replicates is 100% for all internal branches. Labels above branches indicate the level of consensus support (%) based on ML, MP, and NJ.

Results and Discussion

Ortholog Identification

From the seven apicomplexans and the one ciliate examined, we identified 268 single-copy genes that are shared by all eight species. These genes represent less than 10% of the annotated genes from the smallest genome (table 1), indicating that these organisms are highly divergent in their gene content. The long evolutionary distance between ciliates and apicomplexans only partially explains this observation. When the outgroup is not considered, the seven apicomplexans share 508 orthologous genes (of which 433 are single copy in all species). One of our previous studies that examined a different set of apicomplexan species produced similar results and suggested that 28–45% of the genes in an apicomplexan genome are genus-specific (Kuo and Kissinger 2008). This high level of divergence in gene content is consistent with the ancient origin of the phylum. The divergence time between apicomplexans and ciliates was estimated to be in the range of 700–900 Myr based on 129 genes from 36 eukaryotes (Douzery et al. 2004).

For the purpose of phylogenetic analysis, we focus on the 268 single-copy genes shared by all eight species. Many of these genes are responsible for basic cellular processes (e.g., DNA replication, transcription, translation, etc.), as noted in our previous study (Kuo and Kissinger 2008). The sequence identity and annotation information of these genes are provided in supplementary table S1 (Supplementary Material online).

The Apicomplexan Species Tree

The species tree was inferred using two different approaches. The first approach calculated the consensus tree among the 268 individual gene trees, and the second approach utilized a concatenated alignment of 71,830 amino acid sites. Both approaches resulted in the same species tree topology (fig. 1) by all three phylogenetic methods used.

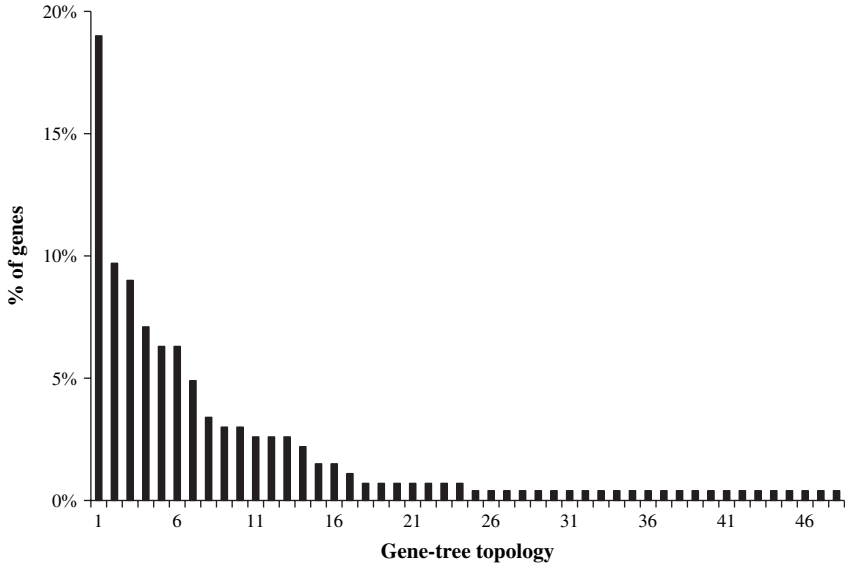

FIG. 2.—Frequency distribution of gene-tree topologies. Based on the 268 single-copy genes examined, we observed a total of 48 gene-tree topologies. The six most frequently observed gene-tree topologies, each supported by more than 5% of the genes, are provided in figure 3.

Groupings of three species pairs, including *P. falciparum* and *P. vivax*, *B. bovis* and *T. annulata*, and *E. tenella* and *T. gondii*, are supported by 87% or more of the genes based on ML consensus. In contrast, the two short internal branches are supported by less than 50% of the genes. Nevertheless, all internal branches received 100% ML bootstrap support based on the analysis of the concatenated alignment.

This tree topology is consistent with most of our prior understanding of apicomplexan evolution based on morphology and development (Perkins et al. 2000), rDNA analyses (Escalante and Ayala 1995; Morrison and Ellis 1997), and multigene phylogenies (Douzery et al. 2004; Philippe et al. 2004; Kuo and Kissinger 2008). The piroplasmids (represented by *B. bovis* and *T. annulata*) form a sister group to the haemosporidians (represented by the *Plasmodium* lineage) with the cyst-forming coccidia (represented by *E. tenella* and *T. gondii*) as the next closely related group. Although the *Cryptosporidium* lineage was classified as a coccidian in early taxonomy work (Levine 1984), our result provides further support to the growing consensus that this lineage is basal to other apicomplexans and separate from other coccidia (Carreno et al. 1999; Zhu et al. 2000; Leander et al. 2003).

The Distribution of Gene Trees

Examination of individual genes revealed a seemingly high degree of incongruence among gene trees. Of the 268 gene trees examined, we observed a total of 48 topologies based on ML analysis (fig. 2). The most frequently observed topology (fig. 3*A*) is consistent with the putative species tree and is supported by 19% of the genes. Each of the next three frequent topologies (fig. 3*B–D*) is supported by approximately 7–10% of the genes and is different in the placement of *C. parvum*. Two additional topologies (fig. 3*E* and *F*) are supported by 6% of the genes and exhibit alternative placements of the *Plasmodium* lineage. The observation that only a relatively small number of topologies are found may be attributed to our limited taxon sampling of eight species. For example, in an analysis of 106 genes from 14 yeast species, Jeffroy et al. (2006) found that each of the genes analyzed supports a distinct topology.

Despite the seemingly high level of incongruence among gene trees, only 16 genes significantly reject the putative species tree topology in the AU test (Shimodaira 2002). When using the more conservative SH test (Shimodaira and Hasegawa 1999), only two genes significantly reject the putative species tree. The first gene is annotated as a hypothetical protein in *P. falciparum* (gene ID: PF14_0326) and exhibits a high level of length variation among the species examined (i.e., varied from 2,452 amino acids in *E. tenella* to 8,094 amino acids in *P. falciparum*). The conserved regions that can be reliably aligned only account for 3% of the alignment. The second gene is annotated as a putative RNA-binding protein in *P. falciparum* (gene ID: PF08_0086) and also exhibits a high level of length variation (i.e., varied from 271 amino acids in *B. bovis* to 1,076 amino acids in *P. vivax*). The protein alignment obtained after GBLOCKS filtering only contains 29 sites. Based on the pattern of sequence length variation, we suspect that the gene annotations may be problematic in some of the species. For this reason, further analysis of these two genes was not pursued.

The finding of a high level of topological incongruence among gene trees that lack statistical significance has been reported in previous genome-scale phylogenetic studies. Lerat et al. (2003) examined 205 single-copy genes shared by 13 gamma-Proteobacteria species and found only two significantly rejected the putative species tree in the SH test. In both cases, the discordance between the gene tree and the putative species tree can be explained by a single lateral gene transfer (LGT) event. Similarly, examinations

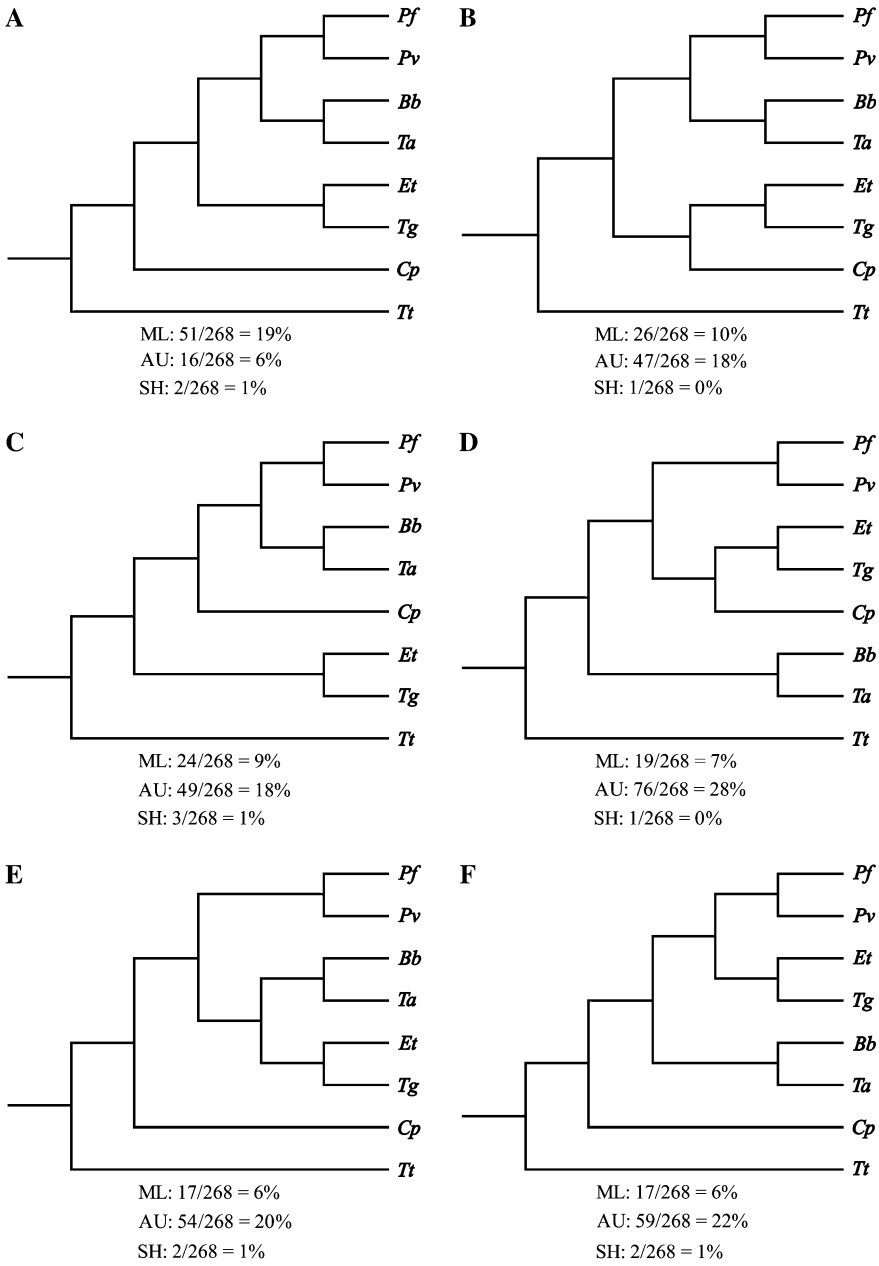

FIG. 3.—The six most frequently observed gene-tree topologies. Each topology is supported by more than 5% of the 268 genes examined. The exact count and frequency of genes that support (or significantly reject) each topology are provided under the tree. ML: frequency of genes that infer the specific topology using ML inference; AU: frequency of genes that significantly reject the topology using AU test; SH: frequency of genes that significantly reject the topology using SH test.

of the 106 single-copy genes shared by a group of *Saccharomyces* spp. showed that the majority of bipartition conflicts among genes have low bootstrap support (Taylor and Piel 2004; Jeffroy et al. 2006).

One possible hypothesis to explain the rare occurrences of a gene significantly rejecting the species tree is that single-copy genes are unlikely to be involved in LGT events (Daubin et al. 2002, 2003). Under this hypothesis, these genes have been confined in the organismal phylogeny throughout their evolutionary history, so the gene-tree topology is unlikely to be radically different from the species tree. By focusing on a small subset of genes that are highly conserved across all apicomplexan lineages examined, our methodology for orthologous gene selection may have effectively excluded genes that experienced LGT since the ciliate–apicomplexan divergence. Although LGT does not appear to influence our phylogenetic inference as presented here, caution should be taken in future studies because several previous studies suggest that LGT is an important evolutionary force in apicomplexans (Huang, Mullapudi, Lancto, et al. 2004; Huang, Mullapudi, Sicheritz-Ponten, and Kissinger 2004; Striepen et al. 2004; Nagamune and Sibley 2006) and other protists (Gogarten 2003; Richards et al. 2003; Andersson 2005).

Table 2
Effects of Removing Genes Based on the Minimum Bootstrap Support

Minimum Bootstrap Cutoff (%)[a]	Number of Genes	Number of Topologies[b]	Percentage of Genes that Inferred	Clade Support Based on ML Consensus (%)	
				((*Pf*, *Pv*), (*Bb*, *Ta*))	(((*Pf*, *Pv*), (*Bb*, *Ta*)), (*Et*, *Tg*))
0	268	48	19	44	38
50	130	25	25	50	40
60	69	15	29	55	49
70	30	10	47	63	60
80	15	5	73	73	80
90	5	1	100	100	100

[a] The bootstrap support for each gene is inferred by the ML method based on 100 replicates. A gene is removed from the analysis if the minimum bootstrap support observed on the gene tree does not meet the cutoff.

[b] Number of observed gene-tree topologies based on ML.

Evaluation of Phylogenetic Signal by Bootstrap Support

To test if the observed topological incongruence among gene trees can be explained by a low resolving power for certain clades in some genes, we used the minimum bootstrap value observed in a gene tree to identify genes that possess strong phylogenetic signals. The results indicate that the percentage of genes that support the putative species tree increases as a function of the bootstrap cutoff used (table 2). In the most extreme example, when only the genes with a minimum bootstrap value of 90% at any node are examined, all five genes that meet this cutoff support the putative species tree topology. Even when the selection stringency is relaxed to a 70% bootstrap support, a cutoff that is commonly used in phylogenetic inference (Hillis and Bull 1993), 47% of these genes are consistent with the putative species tree and the two short internal branches received at least 60% of the consensus support. Curiously, we did not find any significant correlation between bootstrap support and alignment length, average pairwise protein distance, or other attributes of genes (supplementary table S1, Supplementary Material online).

In addition to being consistent with the putative species tree, genes with strong bootstrap support are often insensitive to changes in alignment parameter (table 3), substitution model (table 4), or the phylogenetic method used (table 5). In these tests, we are interested in investigating if a gene could infer the same gene-tree topology across a range of settings used in the phylogenetic inference process; the agreement between the gene-tree topology and the putative species tree is not considered. At 70% minimum bootstrap cutoff, we found that 90% of these genes are robust to a 4-fold change in the gap opening penalty (table 3), 93% of the genes are insensitive to the choice of substitution model (table 4), and 57% of the genes behave consistently across different phylogenetic methods (table 5). Although the use of methodological concordance as a criterion for selecting genes for phylogenetic inference was criticized (Grant and Kluge 2003), our results suggest that a gene is more likely to behave consistently across different phylogenetic methods when it contains a strong phylogenetic signal.

Table 3
Robustness to Alignment Settings as a Function of the Minimum Bootstrap Support

Minimum Bootstrap Cutoff (%)	Percentage of Genes in Each Class[a]		
	Robust[b]	Intermediate[c]	Sensitive[d]
0	60	27	12
50	77	18	5
60	83	16	1
70	90	10	0
80	93	7	0
90	100	0	0

[a] Genes are categorized into three classes based on the sensitivity to sequence alignment settings.

[b] A gene is classified as robust if it produces the same gene-tree topology under all three alignment settings (for details, see Materials and Methods).

[c] A gene is classified as intermediate if it produces the same gene-tree topology under two out of the three alignment settings.

[d] A gene is classified as sensitive if each alignment setting leads to a different gene-tree topology.

Removal of the Long Branches

In addition to the low signal-to-noise ratio in some genes, another possible source of incongruence among gene trees is the LBA problem that resulted from our nonideal taxon sampling. Several observations support this hypothesis. First, when a gene behaved inconsistently across different phylogenetic methods, ML and NJ often result in an identical gene-tree topology that is different from MP (table 5). In addition, the outgroup *T. thermophila* and the ingroup *C. parvum* both have a long evolutionary distance to the other taxa (fig. 1). The lack of additional species that can be used to break up the long branch leading to the *Cryptosporidium* lineage may be responsible for its unstable phylogenetic placement, as evidenced by the fact that three of the most frequently observed gene-tree topologies involve alternative placement of *C. parvum* (fig. 3*B–D*). Although the genome sequence of *C. hominis* is available, adding this species is not particularly helpful. The genomes of these two *Cryptosporidium* spp. exhibit only 3–5% divergence at the nucleotide level (Xu et al. 2004). For the 268 conserved proteins that we used for phylogenetic inference, the sequences from these two species are essentially identical (data not shown).

The issue of nonideal taxon sampling reflects a limitation that is often faced by genome-scale phylogentic inferences (Soltis et al. 2004). To circumvent this limitation, we utilized two other commonly suggested approaches to address the LBA problem (Bergsten 2005). First, all sites that contain gaps or are highly divergent were removed from the

Table 4
Robustness to Substitution Model as a Function of the Minimum Bootstrap Support

	Precentage of Genes in Each Class[a]				
Minimum Bootstrap Cutoff (%)	JTT = LG = WAG	JTT = LG	JTT = WAG	LG = WAG	All Different
0	67	6	10	10	7
50	79	5	7	6	3
60	84	6	4	4	1
70	93	0	0	7	0
80	93	0	0	7	0
90	80	0	0	20	0

[a] Genes are categorized into five classes based on the agreements among the three substitution models used in ML inference. Note that this classification only concerns the consistency of gene-tree topologies inferred by different substitution models for each individual gene. The agreement between a gene tree and the species tree is not considered.

alignment prior to phylogenetic inference by GBLOCKS (see Materials and Methods). Second, we removed either the outgroup *T. thermophila* or the ingroup *C. parvum* prior to sequence alignment and repeated the phylogenetic inference.

When the outgroup is removed from the data set, we observed a large increase in the consensus support for the *Plasmodium–Babesia–Theileria* clade (table 6). Two alternative bipartitions, as shown in panels *E* and *F* of figure 3, received substantially weaker consensus supports regardless of the minimum bootstrap cutoff used. Removal of the ingroup *C. parvum* resulted in a reduction of the number of observed gene-tree topologies (table 6), but the consensus support for the *Plasmodium–Babesia–Theileria* clade is relatively low compared with the removal of *T. thermophila*.

Conclusion

The recent availability of genome sequences allowed us to infer an organismal phylogeny that includes several important apicomplexan pathogens with high confidence. This robust species tree provides a solid foundation for future comparative studies that can improve our understanding of apicomplexan evolution and parasite biology. Although the level of incongruence among gene trees appears to be high at first glance, further investigation indicates that most of the observed conflict does not have strong statistical support. Interestingly, the minimum bootstrap support observed in a gene tree appears to be a useful predictor of phylogenetic performance. Genes that produce strong bootstrap support for all internal branches are more likely to be consistent with the species tree and robust to changes in the alignment parameter or the phylogenetic method used. Nevertheless, examination of multiple unlinked genes with strong phylogenetic signals is important for accurate phylogenetic inference because any single gene can have a different evolutionary history from the organismal phylogeny. Our systematic investigation provides a list of phylogenetically informative genes in the phylum Apicomplexa. These genes are good candidates for future sequencing efforts that aim at improving taxon sampling in this group of important pathogens.

Supplementary Material

Supplementary data files l and 2 and table S1 are available at *Molecular Biology and Evolution* online (http://www.mbe.oxfordjournals.org/).

Acknowledgments

C.-H.K. was supported by a National Institutes of Health (NIH) Training Grant (GM07103), the Kirby and Jan Alton Graduate Fellowship, and a Dissertation Completion Assistantship at the University of Georgia. Funding for this work was provided by NIH R01 AI068908 to J.C.K. P. Brunk, F. Chen, J. Felsenstein, M. Heiges, A. Oliveira, E.

Table 5
Methodological Concordance as a Function of the Minimum Bootstrap Support

	Percentage of Genes in Each Class[a]				
Minimum Bootstrap Cutoff (%)	ML = MP = NJ	ML = MP	ML = NJ	MP = NJ	All Different
0	22	12	25	8	34
50	32	14	34	5	15
60	43	7	38	4	7
70	57	7	33	3	0
80	60	7	33	0	0
90	100	0	0	0	0

[a] Genes are categorized into five classes based on the agreements among the three phylogenetic methods used. Note that this classification only concerns the consistency of gene-tree topologies inferred by different phylogenetic methods for each individual gene. The agreement between a gene tree and the species tree is not considered. Because we used the strict consensus method to consolidate all equally parsimonious trees of a gene, a multifurcating MP tree always has a nonzero topology distance from a fully bifurcating ML or NJ tree.

Table 6
Effects of Taxon Removal

Removal of the outgroup *Tetraymena thermophila* (*Tt*)					
			Consensus support based on ML (%)		
Minimum bootstrap cutoff (%)	Number of genes	Number of topologies	((*Pf*, *Pv*, *Bb*, *Ta*), (*Et*, *Tg*, *Cp*))	((*Pf*, *Pv*, *Et*, *Tg*), (*Bb*, *Ta*, *Cp*))	((*Bb*, *Ta*, *Et*, *Tg*), (*Pf*, *Pv*, *Cp*))
0	268	16	57	20	22
50	215	10	62	17	20
60	169	8	64	16	20
70	124	7	48	15	17
80	81	4	70	14	15
90	42	3	71	10	19
Removal of the ingroup *Cryptosporidium parvum* (*Cp*)					
			Consensus support based on ML (%)		
Minimum bootstrap cutoff (%)	Number of genes	Number of topologies	((*Pf*, *Pv*, *Bb*, *Ta*), (*Et*, *Tg*, *Tt*))	((*Pf*, *Pv*, *Et*, *Tg*), (*Bb*, *Ta*, *Tt*))	((*Bb*, *Ta*, *Et*, *Tg*), (*Pf*, *Pv*, *Tt*))
0	268	18	47	28	23
50	218	12	50	26	23
60	164	10	50	27	23
70	112	5	51	26	23
80	80	5	55	29	16
90	34	3	56	24	21

Robinson, and H. Wang provided valuable assistance on the use of computer hardware and software. We thank the J. Craig Venter Institute for providing prepublication access to the genome sequence data of *P. vivax* and *T. gondii*. The associate editor, Dr Hervé Philippe, and three anonymous reviewers provided constructive comments that greatly improved this manuscript.

Literature Cited

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

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

Andersson JO. 2005. Lateral gene transfer in eukaryotes. Cell Mol Life Sci. 62:1182–1197.

Bahl A, Brunk B, Crabtree J, et al. (18 co-authors). 2003. PlasmoDB: the *Plasmodium* genome resource. A database integrating experimental and computational data. Nucleic Acids Res. 31:212–215.

Bergsten J. 2005. A review of long-branch attraction. Cladistics. 21:163–193.

Brayton KA, Lau AOT, Herndon DR, et al. (28 co-authors). 2007. Genome sequence of *Babesia bovis* and comparative analysis of apicomplexan hemoprotozoa. PLoS Pathog. 3:e148.

Carlton J. 2003. Genome sequencing and comparative genomics of tropical disease pathogens. Cell Microbiol. 5:861–873.

Carreno RA, Matrin DS, Barta JR. 1999. *Cryptosporidium* is more closely related to the gregarines than to coccidia as shown by phylogenetic analysis of apicomplexan parasites inferred using small-subunit ribosomal RNA gene sequences. Parasitol Res. 85:899–904.

Castresana J. 2000. Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. Mol Biol Evol. 17:540–552.

Collins TM, Fedrigo O, Naylor GJP. 2005. Choosing the best genes for the job: the case for stationary genes in genome-scale phylogenetics. Syst Biol. 54:493–500.

Daubin V, Gouy M, Perriere G. 2002. A phylogenomic approach to bacterial phylogeny: evidence of a core of genes sharing a common history. Genome Res. 12:1080–1090.

Daubin V, Moran NA, Ochman H. 2003. Phylogenetics and the cohesion of bacterial genomes. Science. 301:829–832.

Delsuc F, Brinkmann H, Philippe H. 2005. Phylogenomics and the reconstruction of the tree of life. Nat Rev Genet. 6:361–375.

Dopazo H, Dopazo J. 2005. Genome-scale evidence of the nematode-arthropod clade. Genome Biol. 6:R41.

Douzery EJP, Snell EA, Bapteste E, Delsuc F, Philippe H. 2004. The timing of eukaryotic evolution: does a relaxed molecular clock reconcile proteins and fossils? Proc Natl Acad Sci USA. 101:15386–15391.

Eisen JA, Coyne RS, Wu M, et al. (53 co-authors). 2006. Macronuclear genome sequence of the ciliate *Tetrahymena thermophila*, a model eukaryote. PLoS Biol. 4:1620–1642.

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

Felsenstein J. 1989. PHYLIP—phylogeny inference package (version 3.2). Cladistics. 5:164–166.

Gajria B, Bahl A, Brestelli J, et al. (15 co-authors). 2008. ToxoDB: an integrated *Toxoplasma gondii* database resource. Nucleic Acids Res. gkm981. 36:D553–D556.

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.

Gatesy J, Baker RH. 2005. Hidden likelihood support in genomic data: can forty-five wrongs make a right? Syst Biol. 54:483–492.

Gee H. 2003. Evolution: ending incongruence. Nature. 425:782–782.

Gogarten JP. 2003. Gene transfer: gene swapping craze reaches eukaryotes. Curr Biol. 13:R53–R54.

Grant T, Kluge AG. 2003. Data exploration in phylogenetic inference: scientific, heuristic, or neither. Cladistics. 19: 379–418.

Guindon S, Gascuel O. 2003. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. Syst Biol. 52:696–704.
Heiges M, Wang HM, Robinson E, et al. (13 co-authors). 2006. CryptoDB: a *Cryptosporidium* bioinformatics resource update. Nucleic Acids Res. 34:D419–D422.
Hertz-Fowler C, Peacock CS, Wood V, et al. (14 co-authors). 2004. GeneDB: a resource for prokaryotic and eukaryotic organisms. Nucleic Acids Res. 32:D339–D343.
Hillis DM, Bull JJ. 1993. An empirical test of bootstrapping as a method for assessing confidence in phylogenetic analysis. Syst Biol. 42:182–192.
Huang J, Mullapudi N, Lancto CA, Scott M, Abrahamsen MS, Kissinger JC. 2004. *Cryptosporidium parvum*: phylogenomic evidence supports past endosymbiosis, intracellular and horizontal gene transfer. Genome Biol. 5:R88.
Huang JL, Mullapudi N, Sicheritz-Ponten T, Kissinger JC. 2004. A first glimpse into the pattern and scale of gene transfer in the Apicomplexa. Int J Parasitol. 34:265–274.
Huelsenbeck JP, Bull JJ, Cunningham CW. 1996. Combining data in phylogenetic analysis. Trends Ecol Evol. 11:152–158.
Hulsen T, Huynen MA, de Vlieg J, Groenen PMA. 2006. Benchmarking ortholog identification methods using functional genomics data. Genome Biol. 7:R31.
Jeffroy O, Brinkmann H, Delsuc F, Philippe H. 2006. Phylogenomics: the beginning of incongruence? Trends Genet. 22:225–231.
Jones DT, Taylor WR, Thornton JM. 1992. The rapid generation of mutation data matrices from protein sequences. Comput Appl Biosci. 8:275–282.
Kuo C-H, Kissinger JC. 2008. Consistent and contrasting properties of lineage-specific genes in the apicomplexan parasites *Plasmodium* and *Theileria*. BMC Evol Biol. 8:108.
Le SQ, Gascuel O. 2008. An improved general amino acid replacement matrix. Mol Biol Evol. 25:1307–1320.
Leander BS, Harper JT, Keeling PJ. 2003. Molecular phylogeny and surface morphology of marine aseptate gregarines (apicomplexa): *selenidium spp*. and *Lecudina spp*. J Parasitol. 89:1191–1205.
Lerat E, Daubin V, Moran NA. 2003. From gene trees to organismal phylogeny in prokaryotes: the case of the gamma-proteobacteria. PLoS Biol. 1:101–109.
Levine ND. 1984. Taxonomy and review of the coccidian genus *Cryptosporidium* (Protozoa, Apicomplexa). J Protozool. 31:94–98.
Levine ND. 1988. Progress in taxonomy of the Apicomplexan protozoa. J Eukaryot Microbiol. 35:518–520.
Li L, Stoeckert CJ, Roos DS. 2003. OrthoMCL: identification of ortholog groups for eukaryotic genomes. Genome Res. 13: 2178–2189.
Montoya JG, Liesenfeld O. 2004. Toxoplasmosis. Lancet. 363:1965–1976.
Morrison DA, Ellis JT. 1997. Effects of nucleotide sequence alignment on phylogeny estimation: a case study of 18S rDNAs of Apicomplexa. Mol Biol Evol. 14:428–441.
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.
Nishihara H, Okada N, Hasegawa M. 2007. Rooting the eutherian tree: the power and pitfalls of phylogenomics. Genome Biol. 8:R199.
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.
Perkins FO, Barta JR, Clopton RE, Peirce MA, Upton SJ. 2000. Apicomplexa. In: Lee J, Leedale G, Bradbury P, editors. An illustrated guide to the protozoa. Lawrence (KS): Society of Protozoologists. p. 190–369.
Philippe H, Delsuc F, Brinkmann H, Lartillot N. 2005. Phylogenomics. Annu Rev Ecol Evol Syst. 36:541–562.
Philippe H, Lartillot N, Brinkmann H. 2005. Multigene analyses of bilaterian animals corroborate the monophyly of Ecdysozoa, Lophotrochozoa, and Protostomia. Mol Biol Evol. 22:1246–1253.
Philippe H, Snell EA, Bapteste E, Lopez P, Holland PWH, Casane D. 2004. Phylogenomics of eukaryotes: impact of missing data on large alignments. Mol Biol Evol. 21:1740–1752.
Phillips MJ, Delsuc FD, Penny D. 2004. Genome-scale phylogeny and the detection of systematic biases. Mol Biol Evol. 21:1455–1458.
Pollard DA, Iyer VN, Moses AM, Eisen MB. 2006. Widespread discordance of gene trees with species tree in *Drosophila*: evidence for incomplete lineage sorting. PLoS Genet. 2: 1634–1647.
Richards TA, Hirt RP, Williams BAP, Embley TM. 2003. Horizontal gene transfer and the evolution of parasitic protozoa. Protist. 154:17–32.
Robinson DF, Foulds LR. 1981. Comparison of phylogenetic trees. Math Biosci. 53:131–147.
Rodriguez-Ezpeleta N, Brinkmann H, Roure eacute atrice B, Lartillot N, Lang BF, Philippe H. 2007. Detecting and overcoming systematic errors in genome-scale phylogenies. Syst Biol. 56:389–399.
Rokas A. 2006. Genomics and the tree of life. Science. 313:1897–1899.
Rokas A, Williams BL, King N, Carroll SB. 2003. Genome-scale approaches to resolving incongruence in molecular phylogenies. Nature. 425:798–804.
Schmidt HA, Strimmer K, Vingron M, von Haeseler A. 2002. TREE-PUZZLE: maximum likelihood phylogenetic analysis using quartets and parallel computing. Bioinformatics. 18: 502–504.
Shimodaira H. 2002. An approximately unbiased test of phylogenetic tree selection. Syst Biol. 51:492–508.
Shimodaira H, Hasegawa M. 1999. Multiple comparisons of log-likelihoods with applications to phylogenetic inference. Mol Biol Evol. 16:1114–1116.
Shimodaira H, Hasegawa M. 2001. CONSEL: for assessing the confidence of phylogenetic tree selection. Bioinformatics. 17:1246–1247.
Siddall ME. 1995. Another monophyly index: revisiting the jackknife. Cladistics. 11:33–56.
Soltis DE, Albert VA, Savolainen V, et al. (11 co-authors). 2004. Genome-scale data, angiosperm relationships, and 'ending incongruence': a cautionary tale in phylogenetics. Trends Plant Sci. 9:477–483.
Striepen B, Pruijssers AJP, Huang JL, Li C, Gubbels MJ, Umejiego NN, Hedstrom L, Kissinger JC. 2004. Gene transfer in the evolution of parasite nucleotide biosynthesis. Proc Natl Acad Sci USA. 101:3154–3159.
Suyama M, Torrents D, Bork P. 2006. PAL2NAL: robust conversion of protein sequence alignments into the corresponding codon alignments. Nucleic Acids Res. 34:W609–W612.
Tarleton RL, Kissinger J. 2001. Parasite genomics: current status and future prospects. Curr Opin Immunol. 13:395–402.
Taylor DJ, Piel WH. 2004. An assessment of accuracy, error, and conflict with support values from genome-scale phylogenetic data. Mol Biol Evol. 21:1534–1537.
Thompson JD, Higgins DG, Gibson TJ. 1994. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, positions-specific gap

penalties and weight matrix choice. Nucleic Acids Res. 22:4673–4680.
van Dongen S. 2000. Graph clustering by flow simulation. University of Utrecht.
Whelan S, Goldman N. 2001. A general empirical model of protein evolution derived from multiple protein families using a maximum-likelihood approach. Mol Biol Evol. 18: 691–699.
WHO and UNICEF. 2005. World malaria report 2005. Geneva (Switzerland): World Health Organization.
Xu P, Widmer G, Wang Y, et al. (18 co-authors). 2004. The genome of *Cryptosporidium hominis*. Nature. 431:1107–1112.
Zhu G, Keithly JS, Philippe H. 2000. What is the phylogenetic position of *Cryptosporidium*? Int J Syst Evol Microbiol. 50:1673–1681.

Hervé Philippe, Associate Editor

Accepted September 18, 2008