THEORETICAL INVESTIGATION OF HOST DEFENSE STRATEGY EVOLUTION AND PHYLOGENOMIC ANALYSES OF APICOMPLEXAN PARASITES

by

CHIH-HORNG KUO

(Under the Direction of Daniel E.L. Promislow and Jessica C. Kissinger)

ABSTRACT

This dissertation consists of theoretical and phylogenomic approaches to study the evolution of hosts and parasites. The first chapter describes two mathematical models that investigate the evolution of resource allocation strategy in hosts. We found that in a population of hosts that are faced with a potentially costly infection, the evolutionarily stable strategy (ESS) at the host population level is a balanced investment between reproduction and immunity that maintains parasites, even if the host has the capacity to eliminate parasites. Furthermore, hosts exhibiting the ESS can invade any other population through parasite-mediated competition, using the parasites as biological weapons. At the metapopulation level, the dominant strategy is sometimes different from the population-level ESS, depending on the ratio of local extinction rate to colonization rate. This study could help to explain the ubiquity of parasites, and could serve as a framework for investigating parasite-mediated ecological invasions.

The second chapter describes a novel high-throughput method that utilizes genomics data for phylogenetic inference. Results from two exemplar data sets, Vertebrata and Apicomplexa, demonstrate that the identity of phylogenetically informative genes are specific to each taxonomic group, even for phylogenies of similar time scale. The apicomplexans exhibit a high

level of incongruence among gene trees, indicating that a relatively large number of genes are necessary for inferring the species tree. Nonetheless, the availability of genomics data permits the inference of a robust molecular phylogeny that is consistent with our prior knowledge of apicomplexan evolution based on morphology and development.

Using the phylogeny as the foundation, the third chapter is focused on the characterization of lineage-specific (LS) genes in two major apicomplexan lineages, *Plasmodium* and *Theileria*. Consistent with previous studies in animals and bacteria, LS genes have a higher level of sequence divergence in both parasites. The result that many genus- or species-specific genes are putative surface antigens indicates that LS genes could be important in parasite adaptation. The contrasting properties regarding GC content and chromosomal location between LS genes in the two focal genera suggest that closely related parasite lineages can differ in the mechanisms of generating LS genes and their subsequent evolutionary fates.

INDEX WORDS: Evolutionarily stable strategy, Resource allocation, Trade-off, Apicomplexa, Phylogenetic inference, Comparative genomics, Lineage-specific genes

THEORETICAL INVESTIGATION OF HOST DEFENSE STRATEGY EVOLUTION AND PHYLOGENOMIC ANALYSES OF APICOMPLEXAN PARASITES

by

CHIH-HORNG KUO

B.S., National Taiwan University, Taiwan, 1998

M.S., Iowa State University, 2003

A Dissertation Submitted to the Graduate Faculty of The University of Georgia in Partial Fulfillment of the Requirements for the Degree

DOCTOR OF PHILOSOPHY

ATHENS, GEORGIA

2008

THEORETICAL INVESTIGATION OF HOST DEFENSE STRATEGY EVOLUTION AND PHYLOGENOMIC ANALYSES OF APICOMPLEXAN PARASITES

by

CHIH-HORNG KUO

Major Professor:	Daniel E.L. Promislow Jessica C. Kissinger
Committee:	Jeffrey L. Bennetzen Boris Striepen John P. Wares Soojin Yi

Electronic Version Approved:

Maureen Grasso
Dean of the Graduate School
The University of Georgia
May 2008

DEDICATION

To my loving wife Ann Cheng and our wonderful children Alyssa, Vivian, and David.

ACKNOWLEDGEMENTS

I thank my advisors, Daniel Promislow and Jessica Kissinger, for being a team of great mentors. Several other faculty members including John Avise, Jeffrey Bennetzen, Kelly Dyer, Dave Hall, Dale Hoyt, Pejman Rohani, Michael Strand, Boris Striepen, John Wares, and Soojin Yi, served on my dissertation committee and provided helpful comments that shaped my research projects.

My fellow graduate students in the Department of Genetics, especially Vanessa Corby-Harris and Judith Mank, provided invaluable friendship and emotional support that helped me to keep my sanity through graduate school. I am grateful for the generous help and valuable discussions from members of the Promislow lab and the Kissinger lab. The staff members of the department did a great job in making my day-to-day life much easier; Kathy Jones deserves a special thank you in this regard. The computer support provided by Dave Brown, Paul Brunk, Mark Heiges, Adriana Oliveira, Ed Robinson, and Haiming Wang is much appreciated.

For the unconditional love and support that they give me, I owe everything that I achieved to my family. Sadly, my father passed away just months before my graduation and could not see his son earning the first doctorate in our family. Dad, I hope that I have made you proud. I am forever in debt to my parents-in-law, Cheu-Pyeng Cheng and Aina Hung, they treat me as their own and always come to our rescue whenever we need them. Finally, my deepest appreciation goes to my wife Ann Cheng and our children Alyssa, Vivian, and David. They make my life wonderful.

TABLE OF CONTENTS

Page

CHAPTER 1

INTRODUCTION

Motivation and organization of the dissertation

Parasites are ubiquitous in nature and the interactions between parasites and their hosts play an important role in shaping the evolution of both parties involved [1, 2]. Biologists often portray these interactions as episodes of an eternal arms race to highlight the antagonistic and dynamic nature of these interactions. For the parasites, their very survival depends on their ability to invade and exploit the hosts. On the other hand, the selective pressure imposed by the parasites forces the hosts to change almost every aspect of their biology, from immune function to life history strategy [3, 4].

As an evolutionary biologist, I find this perpetual struggle extremely fascinating. More importantly, through my study of pathogens, I also hope to contribute to biomedical research that can improve the quality of life. The three chapters in this dissertation represent my effort to explore the evolution of hosts and parasites through multi-disciplinary approaches. The first chapter is aimed at investigating the evolution of host defense strategy against parasite infections through mathematical modeling. The second and the third chapters integrate phylogenetics and genomics to study genome evolution in an important group of protozoan parasites in the phylum Apicomplexa [5].

Mathematical modeling of host resource allocation strategy

Reproduction and immunity have been the main foci in studies of host life-history evolution because both are important factors in determining the evolutionary fitness of an organism. However, these two traits are not independent of each other since they represent competing needs in the host's overall resource allocation strategy [3]. Mathematical models on this subject have been constructed based on the assumption that there exists a trade-off between reproduction and immunity, a hypothesis that is well supported by empirical evidence [3, 4].

For evolutionary biologists, mathematical models are useful tools for investigating the evolutionary dynamics of a population [6]. In particular, analysis of the evolutionarily stable strategy (ESS) can provide us with insights into the evolutionary trajectory of a trait of interest [7]. Previous studies on the evolution of host resource allocation strategies have found that the ESS for hosts depends on several factors, such as density-dependent regulation of host population density [8], the exact cost of immunity [9-11], the trade-off function between immunity and reproduction [12, 13], the mechanisms of host resistance [14, 15], and the ability of parasites to co-evolve with their host [16].

However, most previous studies were based on the implicit assumption that parasites will always be present in the population. Although a single host can eliminate the parasite by investment in immunity in these models, there is always a non-zero risk of infection. These results lead to the interesting question of why hosts are unable to adopt a strategy that would lead to global extinction of the parasite. To address this question, we have constructed a mathematical model that explicitly allows the hosts to eradicate their parasites from the population [17]. Intriguingly, our result indicates that the ESS for the host is to allow a certain level of infection in the population even when eradicating parasites could lead to a higher population density. This

evolutionary stability is maintained by parasite-mediated competition among hosts; the host exhibiting the ESS is capable of invading other populations by using the parasites as a biological weapon. Furthermore, although parasite-mediated competition plays a major role in the evolution of host strategy, our metapopulation model demonstrated that the dispersal rate and local extinction risk are more important in determining the evolutionary fate above the population level. Taken together, this study may help explain the ubiquity of parasites and can serve as a modeling framework for investigating ecological invasions mediated by parasites.

Apicomplexan parasites as the study system

The protistan phylum Apicomplexa includes approximately 5,000 named species [5], many of which are important pathogens of humans and animals. The most infamous member of this phylum is the causative agent of malaria, *Plasmodium*, which causes more than one million human deaths per year globally [18]. Other important lineages include *Cryptosporidium*, which causes cryptosporidiosis in humans and animals [19, 20], *Theileria*, which causes tropical theileriosis and East Coast fever in cattle [21, 22], and *Toxoplasma*, which causes toxoplasmosis in immunocompromised patients and congenitally infected fetuses [23].

In contrast to bacterial pathogens, these apicomplexan parasites are eukaryotes and share many metabolic pathways with their animal hosts. This fact makes therapeutic target development extremely difficult – a drug that harms an apicomplexan parasite is also likely to harm its host. Currently, there are no effective vaccines or treatments available for most diseases caused by these parasites. Biomedical research on these parasites is challenging because it is often difficult, if not impossible, to maintain live parasite cultures in the laboratory and to genetically manipulate these organisms. Furthermore, the lack of fossil record for these

unicellular organisms has impeded our understanding of how these parasites are related to one another. All this is now changing; the recent release of genome sequences from several apicomplexan species has provided us with new and exciting opportunities to study their biology and evolutionary history [24].

Genome-assisted phylogenetic inference

The second chapter in this dissertation describes a novel high-throughput method for identifying phylogenetically informative genes from a set of genome sequences. We are interested in using this approach to infer the apicomplexan phylogeny because the phylogeny can serve as the foundation for examining evolutionary processes.

Methods developed in the field of molecular phylogenetics have provided biologists with powerful tools to infer phylogeny using sequence data [25]. However, the process still involves several challenges. First, it is often difficult to know which genes one should use *a priori*. Because individual genes can experience horizontal transfer [26], differential losses [27], incomplete lineage sorting [28-30], and other stochastic processes in their evolutionary history, the phylogenetic tree inferred from any given gene may not reflect the species phylogeny. A large-scale analysis of the yeast phylogeny provided a good illustration of this problem [31]. Among the 106 genes examined in this study, the authors were unable to identify characteristics for predicting the phylogenetic performance of individual genes (but see also [32]). Furthermore, four out of six genes that are commonly used for phylogenetic inference produced a gene tree with strong bootstrap support for a topology that is different from the species tree.

In addition to the issue with gene selection, the technical issues associated with phylogenetic analysis *per se* represent another daunting challenge. It is not uncommon to find

that topology of a phylogenetic tree is sensitive to the phylogenetic method used and the associated parameter settings [33-35]. This lack of robustness in phylogenetic inference often creates further complications in the subsequent comparative analyses [36].

Genomic data have been perceived as one solution to these challenges in molecular phylogenetics [37-39], based on the rationale that a large amount of sequence data could provide a sufficiently strong signal to infer the true species tree. Several studies have demonstrated that even difficult phylogenies can be resolved with high confidence by using genomic data [40-42]. Unfortunately, genomic data are only available for a limited number of species to date.

In the case of apicomplexan parasites, genome sequences are available from several species that are important pathogens [24] but not for many other lineages that are important for understanding the evolutionary history of this phylum. The main objective of our study is to identify genes that are useful for phylogenetic inference from available genome sequences. After these phylogenetically informative genes are identified, targeted sequencing effort for these genes from additional species that lack genomic data can be a cost-effective approach to study the apicomplexan phylogeny. We examined two exemplar data sets, including the phylum Apicomplexa and the subphylum Vertebrata, to test the effectiveness of our approach. The results indicate that the identities of informative genes are specific to each taxonomic group, even though the two groups have comparable divergence time [43, 44]. The apicomplexan data set exhibits a low level of congruence among gene trees, suggesting that a relatively large number of genes are necessary to resolve the species tree. Nonetheless, our genome-scale analysis has identified a list of genes that are good candidates for future sequencing efforts to improve taxon sampling.

Evolution of lineage-specific genes

Our genome-scale analysis of apicomplexan phylogeny has inferred a strongly supported species tree that is consistent with our prior understanding of apicomplexan relationship based on morphology and development [45], rDNA analyses [46, 47], and multigene phylogenies [43, 48]. Using this phylogeny as a framework, the third chapter aims to investigate the evolution of lineage-specific genes.

By definition, lineage-specific genes are shared only by a group of closely related organisms [49]. Currently there are several hypotheses regarding the origin of lineage-specific genes, including gene duplication followed by rapid divergence [50, 51], horizontal gene transfer [26, 52-55], intracellular gene transfer between organellar and nuclear genomes [56], exon-shuffling [57, 58], and differential gene loss [59]. The observation that lineage-specific genes often have a higher substitution rate has lead to the hypothesis that they may be important for adaptation and generation of diversity [50, 51, 60]. In the case of parasites, improved knowledge of the lineage-specific genes in these important parasites may lead to a better understanding of their adaptation history and possibly identification of novel therapeutic targets.

Based on the species tree inferred from genomic data, we classified genes in two important apicomplexan pathogens, *Plasmodium falciparum* [18] and *Theileria annulata* [21], into six levels of lineage specificity. In both species,\

the level of sequence divergence is positively correlated with lineage specificity, which provides further support for the hypothesis that gene duplication followed by rapid divergence may be an important mechanism in creating lineage-specific genes. In addition, a large number of genus- and species-specific genes are putative surface antigens that may be involved in host-parasite interaction. This result is consistent with the hypothesis that lineage-specific genes may

be important in adaptation. Interestingly, the two parasite lineages also exhibit several notable differences. The (G + C) content at the third codon position increases with lineage specificity in *P. falciparum* but decreases in *T. annulata*. Furthermore, surface antigens in *Plasmodium* are species-specific and mainly located in sub-telomeric regions. In contrast, surface antigens in *Theileria* are conserved at the genus level and distributed across the entire length of chromosomes. These contrasting properties suggest that the exact mechanisms of generating lineage-specific genes and the subsequent evolutionary fates can differ between related parasite lineages.

References

1. May RM, Anderson RM: **Parasite host coevolution**. *Parasitology* 1990, **100**:S89-S101.

2. Anderson RM, May RM: **Coevolution of hosts and parasites**. *Parasitology* 1982, **85**(OCT):411-426.

3. Sheldon BC, Verhulst S: **Ecological immunology: Costly parasite defences and trade-offs in evolutionary ecology**. *Trends Ecol Evol* 1996, **11**(8):317-321.

4. Zuk M, Stoehr AM: **Immune defense and host life history**. *Am Nat* 2002, **160**:S9-S22.

5. Levine ND: **Progress in Taxonomy of the Apicomplexan Protozoa**. *J Eukaryot Microbiol* 1988, **35**(4):518-520.

6. Otto SP, Day T: **A Biologist's Guide to Mathematical Modeling**. Princeton, N.J.: Princeton University Press; 2007.

7. Maynard Smith J: **Evolution and the Theory of Games**. Cambridge: Cambridge University Press; 1982.

8. Day T, Burns JG: **A consideration of patterns of virulence arising from host-parasite coevolution**. *Evolution* 2003, **57**(3):671-676.

9. Schmid-Hempel P: **Variation in immune defence as a question of evolutionary ecology**. *Proc R Soc Lond B Biol Sci* 2003, **270**:357-366.

10. Lochmiller RL, Deerenberg C: **Trade-offs in evolutionary immunology: just what is the cost of immunity?** *Oikos* 2000, **88**:87-98.

11. Kraaijeveld AR, Ferrari J, Godfray HCJ: **Costs of resistance in insect-parasite and insect-parasitoid interactions**. *Parasitology* 2002, **125**:S71-S82.

12. Boots M, Haraguchi Y: **The evolution of costly resistance in host-parasite systems**. *Am Nat* 1999, **153**:359-370.

13. Bowers RG: **The basic depression ratio of the host: the evolution of host resistance to microparasites**. *Proc R Soc Lond B* 2001, **268**:243-250.

14. van Boven M, Weissing FJ: **The evolutionary economics of immunity**. *Am Nat* 2004, **163**(2):277-294.

15. Boots M, Bowers RG: **Three mechanisms of host resistance to microparasites - Avoidance, recovery and tolerance - Show different evolutionary dynamics**. *J Theor Biol* 1999, **201**(1):13-23.

16. van Baalen M: **Coevolution of recovery ability and virulence**. *Proc R Soc Lond B Biol Sci* 1998, **265**:317-325.

17. Kuo C-H, Corby-Harris V, Promislow DEL: **The unavoidable costs and unexpected benefits of parasitism: Population and metapopulation models of parasite-mediated competition**. *J Theor Biol* 2008, **250**(2):244-256.

18. Gardner MJ, Hall N, Fung E, White O, Berriman M, Hyman RW, Carlton JM, Pain A, Nelson KE, Bowman S *et al*: **Genome sequence of the human malaria parasite *Plasmodium falciparum***. *Nature* 2002, **419**(6906):498-511.

19. Abrahamsen MS, Templeton TJ, Enomoto S, Abrahante JE, Zhu G, Lancto CA, Deng M, Liu C, Widmer G, Tzipori S *et al*: **Complete genome sequence of the Apicomplexan, *Cryptosporidium parvum***. *Science* 2004, **304**:441-445.

20. Xu P, Widmer G, Wang Y, Ozaki LS, Alves JM, Serrano MG, Puiu D, Manque P, Akiyoshi D, Mackey AJ *et al*: **The genome of *Cryptosporidium hominis***. *Nature* 2004, **431**(7012):1107-1112.

21. Pain A, Renauld H, Berriman M, Murphy L, Yeats CA, Weir W, Kerhornou A, Aslett M, Bishop R, Bouchier C *et al*: **Genome of the host-cell transforming parasite *Theileria annulata* compared with *T. parva***. *Science* 2005, **309**(5731):131-133.

22. Gardner MJ, Bishop R, Shah T, de Villiers EP, Carlton JM, Hall N, Ren Q, Paulsen IT, Pain A, Berriman M *et al*: **Genome sequence of *Theileria parva*, a bovine pathogen that transforms lymphocytes**. *Science* 2005, **309**(5731):134-137.

23. Montoya JG, Liesenfeld O: **Toxoplasmosis**. *Lancet* 2004, **363**(9425):1965-1976.

24. Carlton J: **Genome sequencing and comparative genomics of tropical disease pathogens**. *Cell Microbiol* 2003, **5**(12):861-873.

25. Whelan S, Lio P, Goldman N: **Molecular phylogenetics: state-of-the-art methods for looking into the past**. *Trends Genet* 2001, **17**(5):262-272.

26. Gogarten JP, Townsend JP: **Horizontal gene transfer, genome innovation and evolution**. *Nat Rev Microbiol* 2005, **3**(9):679-687.

27. Baldauf SL: **Phylogeny for the faint of heart: a tutorial**. *Trends Genet* 2003, **19**(6):345-351.

28. Maddison WP: **Gene trees in species trees**. *Syst Biol* 1997, **46**(3):523-536.

29. Pamilo P, Nei M: **Relationships between gene trees and species trees**. *Mol Biol Evol* 1988, **5**(5):568-583.

30. Degnan JH, Salter LA: **Gene tree distributions under the coalescent process**. *Evolution* 2005, **59**(1):24-37.

31. Rokas A, Williams BL, King N, Carroll SB: **Genome-scale approaches to resolving incongruence in molecular phylogenies**. *Nature* 2003, **425**(6960):798-804.

32. Collins TM, Fedrigo O, Naylor GJP: **Choosing the best genes for the job: The case for stationary genes in genome-scale phylogenetics**. *Syst Biol* 2005, **54**(3):493-500.

33. Russo CAM, Takezaki N, Nei M: **Efficiencies of different genes and different tree-building methods in recovering a known vertebrate phylogeny**. *Mol Biol Evol* 1996, **13**(3):525-536.

34. Swofford DL, Waddell PJ, Huelsenbeck JP, Foster PG, Lewis PO, Rogers JS: **Bias in phylogenetic estimation and its relevance to the choice between parsimony and likelihood methods**. *Syst Biol* 2001, **50**(4):525-539.

35. Huelsenbeck JP: **Performance of phylogenetic methods in simulation**. *Syst Biol* 1995, **44**(1):17-48.

36. Ronquist F: **Bayesian inference of character evolution**. *Trends Ecol Evol* 2004, **19**(9):475-481.

37. Philippe H, Delsuc F, Brinkmann H, Lartillot N: **Phylogenomics**. *Annu Rev Ecol Evol Syst* 2005, **36**:541-562.

38. Phillips MJ, Delsuc FD, Penny D: **Genome-scale phylogeny and the detection of systematic biases**. *Mol Biol Evol* 2004, **21**(7):1455-1458.

39. Rokas A: **Genomics and the tree of life**. *Science* 2006, **313**(5795):1897-1899.

40. Philippe H, Lartillot N, Brinkmann H: **Multigene analyses of bilaterian animals corroborate the monophyly of Ecdysozoa, Lophotrochozoa, and Protostomia**. *Mol Biol Evol* 2005, **22**(5):1246-1253.

41. Philip GK, Creevey CJ, McInerney JO: **The Opisthokonta and the Ecdysozoa may not be clades: Stronger support for the grouping of plant and animal than for animal and fungi and stronger support for the Coelomata than Ecdysozoa**. *Mol Biol Evol* 2005, **22**(5):1175-1184.

42. Pollard DA, Iyer VN, Moses AM, Eisen MB: **Widespread discordance of gene trees with species tree in *Drosophila*: Evidence for incomplete lineage sorting**. *PLoS Genet* 2006, **2**(10):1634-1647.

43. Douzery EJP, Snell EA, Bapteste E, Delsuc F, Philippe H: **The timing of eukaryotic evolution: Does a relaxed molecular clock reconcile proteins and fossils?** *Proc Natl Acad Sci USA* 2004, **101**(43):15386-15391.

44. Blair JE, Shah P, Hedges SB: **Evolutionary sequence analysis of complete eukaryote genomes**. *BMC Bioinformatics* 2005, **6**.

45. Lee J, Leedale G, Bradbury P: **An Illustrated Guide to the Protozoa**, vol. 1, 2nd edn. Lawrence, KS, USA: Society of Protozoologists; 2000.

46. Escalante A, Ayala F: **Evolutionary origin of *Plasmodium* and other Apicomplexa based on rRNA genes**. *Proc Natl Acad Sci USA* 1995, **92**(13):5793-5797.

47. Morrison DA, Ellis JT: **Effects of nucleotide sequence alignment on phylogeny estimation: A case study of 18S rDNAs of Apicomplexa**. *Mol Biol Evol* 1997, **14**(4):428-441.

48. Philippe H, Snell EA, Bapteste E, Lopez P, Holland PWH, Casane D: **Phylogenomics of eukaryotes: Impact of missing data on large alignments**. *Mol Biol Evol* 2004, **21**(9):1740-1752.

49. Cai J, Woo P, Lau S, Smith D, Yuen K-y: **Accelerated evolutionary rate may be responsible for the emergence of lineage-specific genes in Ascomycota**. *J Mol Evol* 2006, **63**(1):1-11.

50. Domazet-Loso T, Tautz D: **An evolutionary analysis of orphan genes in *Drosophila***. *Genome Res* 2003, **13**(10):2213 - 2219.

51. Alba MM, Castresana J: **Inverse relationship between evolutionary rate and age of mammalian genes**. *Mol Biol Evol* 2005, **22**(3):598-606.

52. Ochman H, Lawrence JG, Groisman EA: **Lateral gene transfer and the nature of bacterial innovation**. *Nature* 2000, **405**(6784):299-304.

53. Lerat E, Daubin V, Ochman H, Moran NA: **Evolutionary origins of genomic repertoires in bacteria**. *PLoS Biol* 2005, **3**:e130.

54. Huang JL, Mullapudi N, Sicheritz-Ponten T, Kissinger JC: **A first glimpse into the pattern and scale of gene transfer in the Apicomplexa**. *Int J Parasitol* 2004, **34**(3):265-274.

55. Hotopp JCD, Clark ME, Oliveira DCSG, Foster JM, Fischer P, Torres MCM, Giebel JD, Kumar N, Ishmael N, Wang S *et al*: **Widespread lateral gene transfer from intracellular bacteria to multicellular eukaryotes**. *Science* 2007, **317**(5845):1753-1756.

56. Huang JL, Mullapudi N, Lancto CA, Scott M, Abrahamsen MS, Kissinger JC: **Phylogenomic evidence supports past endosymbiosis, intracellular and horizontal gene transfer in *Cryptosporidium parvum***. *Genome Biol* 2004, **5**(11).

57. Patthy L: **Genome evolution and the evolution of exon-shuffling -- a review**. *Gene* 1999, **238**(1):103-114.

58. Moran JV, DeBerardinis RJ, Kazazian HH, Jr.: **Exon shuffling by L1 retrotransposition**. *Science* 1999, **283**(5407):1530-1534.

59. Blomme T, Vandepoele K, De Bodt S, Simillion C, Maere S, Van de Peer Y: **The gain and loss of genes during 600 million years of vertebrate evolution**. *Genome Biol* 2006, **7**(5):R43.

60. Wilson GA, Bertrand N, Patel Y, Hughes JB, Feil EJ, Field D: **Orphans as taxonomically restricted and ecologically important genes**. *Microbiology* 2005, **151**(8):2499-2501.

CHAPTER 2

THE UNAVOIDABLE COSTS AND UNEXPECTED BENEFITS OF PARASITISM: POPULATION AND METAPOPULATION MODELS OF PARASITE-MEDIATED COMPETITION[1]

[1] Kuo, C.-H., V. Corby-Harris, and D.E.L. Promislow. 2008. *Journal of Theoretical Biology.* 250:244-256.
Reprinted here with permission of publisher.

Abstract

When faced with limited resources, organisms have to determine how to allocate their resources to maximize fitness. In the presence of parasites, hosts may be selected for their ability to balance between the two competing needs of reproduction and immunity. These decisions can have consequences not only for host fitness, but also for the ability of parasites to persist within the population, and for the competitive dynamics between different host species. We develop two mathematical models to investigate how resource allocation strategies evolve at both population and metapopulation levels. The evolutionarily stable strategy (ESS) at the population level is a balanced investment between reproduction and immunity that maintains parasites, even though the host has the capacity to eliminate parasites. The host exhibiting the ESS can always invade other host populations through parasite-mediated competition, effectively using the parasites as biological weapons. At the metapopulation level, the dominant strategy is sometimes different from the population-level ESS, and depends on the ratio of local extinction rate to host colonization rate. This study may help to explain why parasites are as common as they are, and can serve as a modeling framework for investigating parasite-mediated ecological invasions. Furthermore, this work highlights the possibility that the 'introduction of enemies' process may facilitate species invasion.

Keywords

evolutionarily stable strategy; immunity; invasion; reproduction; resource allocation; trade-off

Introduction

Host-parasite interactions are ubiquitous in nature and are an important force that shapes the life history strategies of both hosts and their parasites [1, 2]. In the presence of parasites, a host's fitness will depend not only on its intrinsic rate of survival and fecundity, but also on its ability to cope with the impact of the parasite. However, these two facets of fitness may be linked to one another. Empirical and theoretical studies suggest that reproduction and immunity represent two competing needs in the host's overall resource allocation strategy [3-5]. While reproduction is obviously essential for population persistence, infections can greatly reduce survival or cause sterility such that resources allocated to reproduction are effectively wasted [6-11]. Because reproduction and immunity are both expensive [12, 13], the way in which a host allocates resources can have a dramatic effect on the host's fitness.

Previous theoretical studies have examined the effect of host resource allocation on the ability to avoid or eliminate infection by parasites. In an optimization model, Medley [14] demonstrated that an individual host can maximize its reproductive value by tolerating some parasite infection. Kaitala [15] considered an evolutionary model in which parasites were able to coevolve with their hosts. In this case, the evolutionarily stable strategy (ESS) [16] for the host population was to maintain phenotypic polymorphisms for immunity. Similarly, van Baalen [17] found that if parasites were not allowed to coevolve with the host, the host's ESS strategy provided some defense, but not enough to eliminate the pathogen altogether.

In these infinite population size models, a single host can eliminate any parasites it carries, but the parasite will always be present in the population, leading to a non-zero risk of infection. Thus, despite the wealth of studies on the ecological and evolutionary dynamics of host-parasite interactions, we do not know how hosts will respond over evolutionary time if

fitness is density dependent, and there exists the possibility of eliminating the pathogen from the population altogether.

When we consider models that assume explicit trade-offs between reproduction and immunity, two general questions arise. First, if hosts can allocate resources to immunity that are sufficient to eliminate the parasite from the population, why are parasites so common? And second, are there potential conflicts between optimal resource allocation strategies at the level of individuals and at the level of groups? If so, how are these conflicts resolved?

Costs of immunity can be separated into two distinct classes—the standing defense cost of maintaining an immune system, and the acute cost of up-regulating the immune system once an individual is infected [13, 18, 19]. While many experimental studies have focused on the acute cost of up-regulating the immune system (reviewed in [13]), artificial selection experiments have demonstrated that costs of standing immunity can lead to decreases in growth, competitive ability, and reproduction [20-22]. These results suggest that a host's optimal strategy for resource allocation will depend on the prevalence and virulence of the parasite in addition to opportunities for reproduction.

Both theoretical models and empirical studies suggest that parasites can significantly alter, and in some cases reverse, the dynamics of both direct [23-26] and indirect competition [27-31]. Furthermore, from an applied perspective, parasites may threaten native biodiversity by affecting the dynamics of interspecific competition and therefore facilitate ecological invasions [32-34]. For example, it is thought that parasite-mediated apparent competition played a role in the ecological displacement of native red squirrels by grey squirrels in England [35]. Additionally, competition experiments conducted in the laboratory demonstrated that a temperate phage could facilitate the invasion of its bacterial host in new environments occupied by other

bacteria [36]. The ubiquity of parasites in nature and their potential for affecting ecological interactions warrants a close look into the mechanisms determining the outcome of parasite-mediated competition.

In this paper, we develop two models to study the evolution of host resource allocation strategy. The first model includes an explicit trade-off function between reproduction and immunity and is used to determine resource allocation strategies that maximize fitness in a population, as well as that which is an evolutionarily stable strategy (ESS), bearing in mind that fitness maximizing strategies and the ESS are not necessarily the same. While a certain resource allocation strategy may allow the hosts to maximize their population density in the presence of parasites, such a strategy may not be an ESS because it leads to a lower competitive ability. Under this scenario, a host population that has a density maximizing strategy may be invaded by other host genotype and be competitively excluded. In contrast to previous models that focus on the cost of mounting an immune response once an individual is infected [e.g., 37], here we focus on the standing costs that are paid by all individuals, regardless of the state of infection. We use evolutionary invasion analysis [38] to determine whether a particular genotype can resist being invaded by all other possible genotypes in a population. We also extend this approach to ask whether being parasitized can affect one's competitive ability relative to a non-parasitized population. Both of our models are constructed based on the interaction between invertebrate hosts and their microbial pathogen. We choose this system for our modeling effort because a large number of empirical studies have greatly improved our understanding of host-parasite interactions [39-41] and because of its increasing importance in evolutionary biology and ecology [42].

We assume in our model that parasites cannot co-evolve in response to changes in the host. Changes in parasite physiology would enhance the ability of the parasite to persist in the population through Red Queen dynamics [43]. By eliminating the possibility of a parasite response, we create a conservative test of the ability of hosts to eliminate pathogens from the population.

Previous models of host-parasite interaction have shown that the evolutionary outcome in a single large population may be quite different from the outcome in a subdivided metapopulation [44-48]. For example, Kirchner and Roy [45] found that in a metapopulation model of host-parasite interaction, selection favored hosts with reduced investment in survival, because in demes where individuals had lower survival, the parasite was less able to persist. Thus, migration rates out of demes where individuals had relatively low survival rates were actually higher. In a model with explicit trade-offs between allocation to immunity and reproduction, we might expect the optimal allocation strategy to shift under the conditions of a subdivided metapopulation. In the second part of this study, we explore whether a metapopulation structure alters the patterns that we observe in the within-population model. While the models presented here are relatively simple, they offer an important step towards furthering our understanding of the trade-offs that hosts face in deciding how best to cope with the ever-present risk of parasitism.

Model I. Resource allocation strategies — A single-population model

Model description

This model is modified from one developed by Kirchner and Roy [45], which explores the role of parasites in the evolution of host lifespan. To investigate the evolutionary dynamics of

resource allocation strategies, we have changed their original susceptible-infected (*SI*) model into a susceptible-infected-susceptible (*SIS*) model (Figure 2.1), which allows infected individuals to recover. We introduce a resource-dependent trade-off between recovery and reproduction, and then determine how hosts allocate resources to these two traits. We assume a linear trade-off function between the amount of resources available to reproduction versus immunity, as in the classic Y-model for resource allocation [49]. The host-parasite interaction model that we discuss here focuses on the population dynamics of the host rather than the parasite, as we are interested in the evolution of the host's resource allocation strategy. All hosts are in the susceptible class (*S*) at birth, and become infected via horizontal transfer from already infected hosts. Hosts recover from the infection at a rate that is positively correlated with the host's investment in immunity. However, a host genotype that maintains a high immunity level will have a low reproductive rate, and hence, low fitness.

We limit our analysis to trade-offs between reproduction and immunity that are due to the evolutionary cost of maintaining a high immunity level, rather than the cost of mounting an immune response after infection [18]. Such maintenance costs are apparent in the cellular defense response in insects. Insects with a relatively high concentration of circulating hemocytes are more effective at clearing invading parasites at the onset of an infection [50]. However, this immune capability may come at the cost of diverting resources from other fitness components. For example, in a selection experiment carried out with a population of *Drosophila melanogaster* and its natural parasitoid, *Asobara tabida*, parasitoid-resistant lines doubled the number of circulating haemocytes but had a low larval competitive ability [21, 51].

In our model, recovered hosts move from the infected class (*I*) back to the susceptible class and are not immune to future infections. This model is appropriate for invertebrate systems

in which the hosts have innate immunity but not acquired immunity (but see [52]). We express the model as a pair of ordinary differential equations as follows:

$$\frac{dS}{dt} = rx(1 - S - I)S - \beta SI + r(1 - x)I - \mu S \quad (1)$$

$$\frac{dI}{dt} = \beta SI - r(1 - x)I - m\mu I \quad (2)$$

where the density of $S + I \leq 1$. As in standard SIS models, β is the parasite transmission rate, infection occurs at a rate βSI, μ is the mortality rate of the host, and m is the ratio by which infection increases host mortality rate ($m \geq 1$) and can be viewed as a measurement of parasite virulence. Thus, susceptible hosts die at a rate μS and infected hosts die at a rate $m\mu I$. We chose to use a multiplicative form of infection mortality instead of the more standard additive form to account for the possible interactions between different mortality sources [53]. A more general model that partitioned the mortality rate into extrinsic and intrinsic components and linked the intrinsic mortality rate to the resource allocation term did not produce qualitatively different predictions (results not shown).

To model resource allocation, we include a term r, which is the resource acquisition rate of the host, and x, which is the proportion of resource allocated to reproduction ($0 < x \leq 1$) (with $1 - x$ equal to the proportion allocated to recovery from infection). The potential per capita reproduction rate in the absence of carrying capacity constraints is given by rx and the per capita recovery rate from infection is simply $r(1 - x)$. We assume that host population growth is under density-dependent regulation, with the term $(1 - S - I)$ representing the unoccupied fraction of carrying capacity that is available for new individuals. The host population can never reach a density of one in this model due to its own mortality. To simplify the model, we assume that infected individuals are sterile, but regain full fertility once they have recovered.

Anderson and May [54] define the basic reproductive ratio of the parasite (R_0) as the average number of secondary infections produced when one infected individual is introduced into a host population where everyone is susceptible. We can calculate R_0 by multiplying the transmission rate by the duration of the infection period and host density:

$$R_0 = \frac{\beta}{\left(r(1-x)+m\mu\right)}\left(1-\frac{\mu}{rx}\right) \tag{3}$$

Three equilibria (host extinction, parasite extinction, and host-parasite coexistence) exist in this system. The trivial equilibrium occurs when the host reproduction rate is less than the host mortality rate:

$$S^* = 0 \text{ and } I^* = 0 \text{ if } rx < \mu \tag{4}$$

The host population can persist and eliminate parasites when two conditions are met. First, the host reproduction rate is higher than the host mortality rate. Second, the basic reproductive ratio of the parasite is less than one. In other words, the parasite cannot persist if each infected host recovers or dies from an infection before spreading the parasite to at least one susceptible host:

$$S^* = 1-\frac{\mu}{rx} \text{ and } I^* = 0 \text{ if } rx > \mu \text{ and } R_0 < 1 \tag{5}$$

When parasites are absent, the equilibrium host density increases with investment in reproduction (x), as shown in Figure 2.2A. When reproductive rates of hosts and parasites are sufficiently high, stable coexistence of hosts and parasites can occur, with equilibrium values of infected and uninfected individuals given by:

$$S^* = \frac{m\mu + r(1-x)}{\beta} \text{ and } I^* = \frac{(m\mu + r - rx)[\beta(rx-\mu) - rx(m\mu + r - rx)]}{\beta[\beta m\mu + rx(m\mu + r - rx)]}$$

$$\text{if} \quad rx > \mu \text{ and } R_0 > 1 \tag{6}$$

In the presence of parasites, the equilibrium density of susceptible hosts increases with the host recovery rate $r(1 - x)$. When the host recovery rate is sufficiently high, an infected host either recovers or dies before infecting another host and eventually drives the parasite to extinction (Figure 2.2B).

The density maximizing strategy and the evolutionarily stable strategy

We define the fitness of a host genotype as the population density at equilibrium (i.e., S^* in the absence of parasites or $S^* + I^*$ in the presence of parasites). The rationale behind this definition is that the total density of hosts at equilibrium can be used as an indicator of relative colonization rate of a genotype in a metapopulation context, assuming a small proportion of hosts migrate into other patches in each generation. This definition of fitness is relevant in our second model concerning resource allocation strategy evolution at the metapopulation level.

In the absence of parasites, the density of hosts increases with investment in reproduction (eq [5], see Figure 2.2A for an example). The resource allocation strategy that maximizes the host density is simply to invest fully in reproduction (i.e., $x = 1$).

Based on our model assumption that the hosts can increase recovery rate through heavier investment in immunity, a strategy that eliminates the parasites and confers resistance to future epidemics exists in most of the parameter space (see the peak of S* in Figure 2.2B for an example). Under this condition, an infected host always recovers or dies before spreading the parasite to another susceptible host such that the parasite can never spread (i.e., $I^* = 0$). This immunizing strategy (x_{imu}) is derived by solving for the stability boundary that separates the equilibria for parasite extinction (eq [5]) and host-parasite coexistence (eq [6]), and is given by:

$$x_{imu} = \frac{(r + m\mu - \beta) + \sqrt{(r + m\mu - \beta)^2 + 4\beta\mu}}{2r} \qquad (7)$$

The immunizing strategy (x_{imu}) may or may not be a strategy that maximizes the total density of hosts at equilibrium. When conditions favor the coexistence of hosts and parasites, the strategy that maximizes the total density of hosts is derived by solving for the maximum of ($S^* + I^*$) in eq (6). This density maximizing strategy, x_{max}, and its condition boundary are given by:

$$x_{max} = \frac{-\beta m + (m-1)(r+m\mu) + \sqrt{\beta m(\beta m - (m-1)(r-\mu+2m\mu))}}{r(m-1)}$$

if $$\beta m - (m-1)(r-\mu+2m\mu) \geq 0 \qquad (8)$$

When conditions favor the extinction of parasites, the density maximizing strategy is equivalent to the strategy that eliminates the parasites (i.e., $x_{max} = x_{imu}$).

In the context of within-population evolution, we are interested in identifying the evolutionarily stable strategy (ESS) that prevents the invasion of other genotypes. The ESS is a strategy that ensures the persistence of a particular genotype, and may or may not be a strategy that eliminates the parasites or maximizes host density. This distinction is important because when the strategy that maximizes host density is not evolutionarily stable (i.e., the x_{max} genotype can be invaded by other genotypes) it is generally considered irrelevant from an evolutionary perspective. The ESS for resource allocation (x_{ESS}) is analytically derived by expanding the model to include multiple host genotypes and examining the Jacobian matrix (see Appendix A). The resulting ESS is:

$$x_{ESS} = \frac{1}{2(-1+2m)r} \times \left(\begin{array}{l} (-2+3m)r - 2m(\beta+\mu) + 3m^2\mu \\ + \sqrt{m}\sqrt{4\beta(r-\mu) + 2m^2(r-2\beta)\mu + m^3\mu^2 + m\left[r^2 + 4\beta(\beta+3\mu-r)\right]} \end{array} \right) \qquad (9)$$

To better understand the implications of these analytical solutions, we numerically compare x_{imu}, x_{max} and x_{ESS} under a range of parameter settings. Figure 2.3 illustrates how each of

the four model parameters (i.e., r, μ, β, and m) affects the values of x_{imu}, x_{max} and x_{ESS}. Under all conditions, the ESS requires the same or a higher investment in reproduction than the other strategies (i.e., $x_{ESS} \geq x_{max} \geq x_{imu}$).

When the resource acquisition rate is extremely low ($r < 0.6$ in Figure 2.3A), investment in reproduction is fundamental for survival and x_{ESS} is equal to one. As resource availability increases, the ESS becomes a balance between reproduction and immunity but is never a strategy that eliminates the parasite. Compared to the ESS, the density maximizing strategy requires a higher investment in immunity and the immunizing strategy requires the highest investment in immunity. When the resource availability becomes extremely high, the host only needs to invest a small fraction of resource acquired in immunity to eliminate the parasite and maximize population density. Under this condition, x_{max} and x_{imu} become equivalent and converge toward x_{ESS}. Both x_{ESS} and x_{imu} monotonically increase with the extrinsic mortality rate as reproduction becomes more important under a higher mortality rate (Figure 2.3B). Parasite extinction occurs when the host mortality rate is sufficiently high (eq[5]) and all three strategies lead to full investment in reproduction (i.e., $x_{ESS} = x_{max} = x_{imu} = 1$). Because the infection prevalence increases with the parasite transmission rate, a high investment in immunity (i.e., a lower x) would be required for the host to eliminate the parasite (Figure 2.3C). In contrast, the immunity investment required to eliminate the parasite decreases as virulence increases (Figure 2.3D). When the virulence becomes too high, the parasite kills the host before it can spread to another susceptible individual and the infection cannot sustain itself in the host population. All three strategies become full investment in reproduction under this situation.

Pairwise invasibility plots [55] are useful tools for visualizing the evolutionary trajectory of the resource allocation strategy in a population. In the absence of the parasite, the host evolves

toward full investment in reproduction (i.e., $x_{ESS} = 1$, see Figure 2.4A). With few exceptions (see discussion of Figure 2.3, above), the ESS in the presence of the parasite is a strategy that balances reproduction and immunity (Figure 2.4B). The host genotype with this evolutionarily stable resource allocation strategy is resistant to invasions and can invade populations of all other genotypes when parasites are present. Intriguingly, evolutionary stability is maintained by parasite-mediated competition and density-dependent regulation. When the resident host has a high level of investment in reproduction and low immunity (i.e., $x_{ESS} < x_R$, see the right hand part of Figure 2.4B), infection is prevalent in the population (see the right hand part of Figure 2.2B). Under this scenario, the resources invested in reproduction are mostly wasted because the host cannot reproduce during the infection period. A mutant can invade the population if it can recover faster from infection and reproduce, provided that its investment in reproduction is sufficiently high. The result is that the host evolves toward a lower investment in reproduction (i.e., a smaller x) until it reaches the ESS. In contrast, when the resident host has a low level of investment in reproduction and high immunity ($x_R < x_{ESS}$, see the left hand side of Figure 2.4B), parasites are rare and the risk of infection is low. A successful mutant would require a higher investment in reproduction than the resident genotype while maintaining a sufficiently high recovery rate. Under this scenario, the host evolves toward a higher investment in reproduction (i.e., a larger x) until it reaches the ESS. Thus, regardless of the initial condition, the host always evolves toward a single ESS under this model.

The predictions from our population model have several implications. First, in most of the parameter space, the host actually has the potential to eliminate the parasite through high investment in immunity (i.e., by adapting x_{imu}), assuming that the parasite does not evolve some counter-adaptation. However, such a strategy is not evolutionarily attainable because natural

selection acts at the individual level, rather than the population level. Instead of eliminating parasites, the stable equilibrium state is host-parasite coexistence. Second, in the presence of parasites, a single ESS exists and the host evolves toward this ESS regardless of the initial condition. Furthermore, the host with the x_{ESS} genotype can invade populations of any other genotype and is resistant to counter-invasion.

Model II. Resource allocation strategies —A metapopulation model

In our second model, we extend our earlier results to determine the evolutionary dynamics in a metapopulation using the framework developed by Nee and May [56]. The metapopulation consists of three types of demes, including those in which the parasite is absent (*H,* for "host-only", occupied by hosts with genotype $x_H = 1$), those in which the parasite is present (*P*, for "parasite-present", occupied by hosts with genotype $x_p = x_{ESS}$), and extinct demes in which there is no host (*V*, for "vacant").

Our metapopulation model rests on four assumptions. First, the dynamics at the population level occur at a much faster time scale compared to the metapopulation dynamics, such that host population density is at equilibrium in all local patches immediately following colonization or invasion. Based on our analysis of within population dynamics (see Figure 2.5 for an example), the density of hosts reach equilibrium in a few generations after a colonization or invasion event. Assuming that local extinction is a relatively rare event (i.e., the expected time to local extinction is longer than the time for a host population to reach equilibrium density), this approximation using the separation of time scales greatly simplifies the model and allows us to track the dynamics analytically. Second, we assume that all local patches are identical (i.e., r, μ, β, and m are the same in all demes). Under certain conditions (e.g., low resource availability,

high extrinsic mortality, or high parasite virulence, see Figure 2.3), x_{ESS} is equal to one (i.e., $x_P = x_H = 1$) regardless of whether the parasite is present in the local patch, such that all hosts have the same genotype. To obtain biologically meaningful results, we focus only on the condition where the x_{ESS} is less than one in the presence of parasites (i.e., $x_P < x_H = 1$). Third, the colonization rate of a genotype is a monotonically increasing function of its population density. And fourth, parasites can only disperse with their hosts and always accompany the x_P host in a colonization event. In other words, we assume all the susceptible x_p hosts carry dormant parasites and the infection status has no effect on dispersal. Examples of dormant parasites that can be transmitted by susceptible hosts include several bacterial and fungal pathogens of insects [57, 58].

Based on our findings in Model I, the three patches in this model can be described as: (1) patches that are occupied by hosts but not parasites, in which the hosts evolve to become inferior competitors but good colonizers. These patches occur with a frequency F_H, (2) patches occurring with frequency F_P that are occupied by hosts and parasites, where the hosts evolve to become superior competitors but poor colonizers; and (3) vacant patches with frequency F_V that are a result of a local extinction event and are available for colonization immediately. This model does not consider the spatial structure of metapopulations, but could be extended to investigate spatial dynamics using cellular automata techniques.

Figure 2.6 is a schematic representation of this metapopulation model, which can be expressed as a system of ordinary differential equations:

$$\frac{dF_H}{dt} = c_H F_H F_V - c_P F_H F_P - eF_H \tag{10}$$

$$\frac{dF_P}{dt} = c_P F_P F_V + c_P F_H F_P - eF_P \tag{11}$$

$$\frac{dF_V}{dt} = eF_H + eF_P - c_H F_H F_V - c_P F_P F_V \tag{12}$$

where c is the colonization rate of each host genotype, e is the local extinction rate and $F_H + F_P + F_V = 1$. Based on our second and third assumptions, we assume that $c_P < c_H$.

The equilibrium frequencies for the three patch types are affected by the ratio of local extinction rate to the host colonization rate (Figure 2.7). Global extinction occurs when the local extinction rate is higher than the colonization rate of the x_H genotype:

$$F_H^* = 0,\ F_P^* = 0, \text{ and } F_V^* = 1 \text{ if } c_P < c_H < e \tag{13}$$

Host persistence and parasite extinction occur when the local extinction rate takes a value between the colonization rates of the two genotypes:

$$F_H^* = 1 - \frac{e}{c_H},\ F_P^* = 0, \text{ and } F_V^* = \frac{e}{c_H} \text{ if } c_P < e < c_H \tag{14}$$

The two host genotypes can coexist when the local extinction rate is lower than the colonization rate of the x_P genotype and is higher than the ratio of the colonization rate of the x_H genotype to that of the x_P genotype. Under this condition, parasites coexist with their host in some local patches (F_P^*) but not others (F_H^*); a higher local extinction rate reduces the parasite prevalence at the metapopulation level (i.e., F_P^* decreases when e increases):

$$F_H^* = \frac{e}{c_P} - \frac{c_P}{c_H},\ F_P^* = 1 - \frac{e}{c_P}, \text{ and } F_V^* = \frac{c_P}{c_H} \text{ if } \frac{c_P}{c_H} < \frac{e}{c_P} < 1 \tag{15}$$

When the local extinction rate is too low, parasites can spread to all non-vacant patches and eliminate the host with the x_H genotype:

$$F_H^* = 0,\ F_P^* = 1 - \frac{e}{c_P}, \text{ and } F_V^* = \frac{e}{c_P} \text{ if } \frac{e}{c_P} < \frac{c_P}{c_H} \tag{16}$$

Figure 2.8 shows the domain of feasibility of this metapopulation model.

Discussion

Ever since the groundbreaking work of Anderson and May [59], a quarter century of theoretical and empirical work on host-parasite interactions has taught us much about host population dynamics when faced with parasites. However, until recently, few models have considered that the resources that a host uses to fight off parasites may come at the expense of investment in other fitness-related traits. We have used several approaches here to explore the consequences of these trade-offs, both in terms of the fitness of the host, and of the ability of the parasite to persist in the population.

Two important points are illustrated by our results. First, while sufficient investment in immune function may enable a population to eliminate a parasite altogether, selection at the individual level makes this an evolutionarily unstable strategy. Second, an infected genotype with an ESS level of investment can always invade an uninfected population. This suggests that carriers of a parasite can use the parasite as a Trojan horse, allowing it to invade a healthy population. We discuss both of these issues in greater detail below.

Within-population models of host-parasite dynamics

In a relatively simple model with trade-offs between host reproductive rate and the rate of recovery from infection, we found that it was possible for the host population to eliminate the parasites in certain parameter space by evolving a high recovery rate. This assumes, of course, that the parasite has not evolved a counter-measure to overcome the strong defenses of the host. Nonetheless, our results indicate that the ESS for the host is to maintain a certain level of infection within the population even when it is theoretically possible to eliminate the parasites.

This ESS is an evolutionary trap and does not depend on the initial condition of the host population.

The results of our model rest on four particular assumptions. First, as with many previous models [e.g., 60, 61-63], we assumed that the population size was regulated by density-dependent reproduction. Day and Burns [37] showed that density-dependent regulation results in an intermediate ESS for immunity, whereas the population evolves either toward maximal or minimal immunity when there is no density-dependent regulation. If no single ESS exists or the direction of evolution depends on the initial condition of the population, the evolutionary dynamics at the metapopulation level can be substantially more complex.

Second, we also assumed that the trade-off between investment in reproduction and investment in immune function was linear. An obvious benefit of this assumption is that it makes the model analytically tractable. However, other shapes for the trade-off (e.g., concave, convex, sigmoidal) have been shown previously to alter the ESS for susceptibility [63, 64]. For example, in an SIS model with a trade-off between rates of resource acquisition (r) and parasite transmission (β), Boots and Haraguchi [64] found that if the trade-off function were linear or convex, the parasite strain with the minimum β would always win. In contrast, with a concave trade-off function, strains with either a minimum or maximum β could win.

Third, we assume that organisms trade off reproductive capacity for recovery rate, and that the cost is paid independent of whether or not a host is infected (i.e., hosts incur a constitutive standing cost of immunity, as opposed to a facultative response cost). This assumption is consistent with recent studies showing that fitness consequences of exposure to parasites may be due in part to the cost of immunity rather than the cost of infection [5, 65]. Recent models have shown that the evolutionary dynamics of host-parasite models may differ

depending on whether hosts try to fight off parasites through lower transmission rates, higher recovery rates, or higher tolerance [60, 61]. Models have also explored the cost of maintaining the baseline level of immunity, versus the cost that is incurred when genes are up-regulated in response to a parasite [13, 18, 19]. The costs of up-regulating defense machinery are only incurred by the infected individuals, whereas the standing costs for maintaining the immune system are incurred by both susceptible and infected individuals. Here, too, the evolutionary dynamics of host-parasite interactions depend on the specific cost scenario and immunity component involved (i.e., enhanced recovery, reduced virulence, and reduced susceptibility) [60].

Under the simplifying assumptions inherent in our model, we found that the host ESS did not eliminate the pathogen from the population. Future studies should consider whether a similar strategy would evolve under different types of defense against parasites, and for standing versus facultative costs.

Finally, we have focused on the way in which hosts evolve in the presence of parasites, assuming that parasite transmission rate and virulence remain constant for the time scale concerned in this model. Our original rationale for doing so was to create a conservative test of the hypothesis that hosts would not evolve to eliminate parasites from a population, even when it was biologically possible. By preventing parasites from co-evolving in response to host change, it should be easier for hosts to drive the parasite to extinction. As shown in Figure 2.3, an immunizing strategy exists for the host in all parameter space examined. However, the immunizing strategy for eliminating the parasites always requires more investment in immunity than the ESS and thus is not attainable by the host from an evolutionary perspective.

There may be some biological justification for the assumption of constant parasite transmission rate and virulence as well. First, parasites with little or no host-specificity may not have a strong response to the life history evolution of one particular host species. Second, parasite may evolve in response to the change of host strategy but the coevolutionary dynamic keeps these two parameters relatively stable. Artificial selection experiments demonstrate that the evolution of host resource allocation strategy can occur very quickly, reaching a maximal response in as little as five generations [21]. This empirical result provides justification for our simplifying assumption that transmission rate and virulence do not change in response to host evolution. In the more complex situation in which parasite transmission rate and virulence evolve in response to changes in host resource strategy, more than one evolutionarily stable outcome may be possible. A coevolution model by van Baalen [17] found that the ESS for the hosts is to have a limited investment in recovery rate when parasites do not coevolve. However, two possible outcomes exist when the parasites coevolve to adapt to their hosts. In the first scenario, parasites become relatively avirulent and the host invests little in recovery ability. The second scenario corresponds to an escalated arms race in which the host invests heavily in recovery ability to defend against rare but virulent parasites [17].

Invasion analysis

Our invasion analysis demonstrated that the ESS genotype could displace all other genotypes in the presence of the parasites, even when the resident genotype has a high level of investment in immunity that is sufficient to eliminate the parasites in isolation. Under this scenario, the invading hosts act as reservoirs such that the parasites are always present in the population. Even though the resident hosts can recover from an infection at a faster rate than the

invaders, the invaders are superior competitors because of their higher reproduction rate. This result may inform our understanding of the role that parasites play in the process of invasion.

The conventional 'escape from enemies' hypothesis has provided a popular explanation for the success of invading species. The idea is that an invading species that leaves its costly parasites behind can out-compete local species, which must reserve some of their resources for immunity. This model predicts that invaders should suffer lower levels of damage from parasites relative to the local species, should evolve greater investment in reproduction relative to the populations from which they originate, and should be less able to fight off parasites when reintroduced to their native habitat. Some studies have found strong support for this hypothesis [66-68].

But one might equally turn this idea on its head, arguing that invaders could succeed by introducing enemies to which resident populations are not adapted. While few explicit models of this idea exist [35, 36], one can trace the genesis of the idea to early work on apparent competition. In the first model of apparent competition in a host-microparasite system, Holt and Pickering [28] found that in the presence of a parasite that infects two distinct populations, the population with relatively low resistance would be displaced by the more resistant population. In the model we present here, we show the more general result that any non-parasitized population can be displaced, even if it has higher resistance than an invading parasitized population. This result has direct applications to understanding the causes and dynamics of ecological invasions. Our result is similar to an idea developed by Wodarz and Sasaki [69], who suggest that a parasite can be used as a biological weapon in inter-specific competition, leading to the evolution of suboptimal immunity in the host. We note, however, that their results are based on the assumption that cross-specific infection is always lethal, which is not likely to be generally true.

Metapopulation models of host-parasite dynamics

It is now well-established that solutions for evolutionarily stable strategies within populations do not necessarily predict the outcome of selection acting on a metapopulation [70, 71]. For example, metapopulation models have demonstrated that in subdivided populations, life history traits can evolve not simply due to their direct correlation with fitness, but also because of their effect on rates of extinction and colonization [71]. Of relevance to the models discussed here, a metapopulation structure can increase the ability of competitors to coexist [71].

We showed that for a single panmictic population, parasites create an evolutionary trap, in which a strategy that eliminates the parasite is never favored by selection. This is not necessarily the case, however, in a metapopulation model, in which the ESS at the population level may or may not be the dominant strategy at the metapopulation level. Because the competitive advantage of the ESS genotype over the full reproduction genotype stems from parasite-mediated competition, the parasite prevalence at the metapopulation level determines the success of the ESS genotype. When local extinction rate is high compared to host colonization rate, parasites are restricted to a small fraction of local patches and the dominant strategy is to fully invest in reproduction. These results are in a similar vein to work by Kirchner and Roy [45], who found that in a metapopulation, parasites could favor the evolution of reduced lifespan of hosts. If hosts died before they had a chance to transmit the parasite to other individuals within a population, migration rates to other populations would increase.

The metapopulation model that we present here makes no assumptions about underlying spatial structure. However, especially in the context of parasite transmission, spatial structure may have important consequences for host-parasite dynamics [72-75]. In particular, the structure of the host contact network within which the parasite is transmitted can affect the dynamics of epidemics [76-81]. In addition to studies of the effect of network structure on host infection

rates, others have studied how network structure can affect the evolution of parasite virulence [82]. However, the effects of host contact network structure on the coevolutionary dynamics of both host and parasites have yet to be investigated.

Conclusion

The ecology and evolution of the invertebrate immunity has attracted much research attention in recent years [42]. In addition to studies that have focused on the underlying physiological and genetic mechanisms of invertebrate immune system [39-41], mathematical models have proved indispensable in developing explicit hypotheses for empirical studies. Through the use of relatively simple SIS and metapopulation models, here we show that trade-offs between reproduction and immunity may be able to explain why certain populations or species are able to invade others, and why parasites remain as common as they do, despite the evolution of elaborate host defenses to ward them off. While these simple models have allowed us to draw some rather general conclusions, we are now in need of more detailed models to determine the validity of our simplifying assumptions, and of further empirical studies to test the claims about host-parasite coevolution and host invasions that have been put forward here and elsewhere.

Acknowledgements

We thank Troy Day, Susan Elliott, Pejman Rohani, members of the Promislow lab, and three anonymous reviewers for helpful comments. Troy Day and Ted Shifrin provided assistance on the ESS analysis. CHK was supported by a National Institutes of Health Training Grant (GM07103), a Kirby and Jan Alton Graduate Fellowship, and a Dissertation Completion

Assistantship at the University of Georgia. VCH was supported by a Dissertation Improvement Grant from the National Science Foundation (DEB-0508785) and a Kirby and Jan Alton Graduate Fellowship in Genetics. Funding for this work was also provided by an Ellison Medical Foundation Senior Scholar Award to DELP.

References

1. Anderson RM, May RM: **Coevolution of hosts and parasites**. *Parasitology* 1982, **85**(OCT):411-426.

2. May RM, Anderson RM: **Parasite host coevolution**. *Parasitology* 1990, **100**:S89-S101.

3. Sheldon BC, Verhulst S: **Ecological immunology: Costly parasite defences and trade-offs in evolutionary ecology**. *Trends Ecol Evol* 1996, **11**(8):317-321.

4. Zuk M, Stoehr AM: **Immune defense and host life history**. *Am Nat* 2002, **160**:S9-S22.

5. Zerofsky M, Harel E, Silverman N, Tatar M: **Aging of the innate immune response in *Drosophila melanogaster***. *Aging Cell* 2005, **4**(2):103-108.

6. Baudoin M: **Host castration as a parasitic strategy**. *Evolution* 1975, **29**:335-352.

7. Clay K: **Parasitic castration of plants by fungi**. *Trends Ecol Evol* 1991, **6**:162-166.

8. Lively CM: **Evidence from a New Zealand snail for the maintenance of sex by parasitism**. *Nature* 1987, **328**:519-521.

9. Ebert D, Carius HJ, Little T, Decaestecker E: **The evolution of virulence when parasites cause host castration and gigantism**. *Am Nat* 2004, **164**(5):S19-S32.

10. Roy BA: **Floral mimicry by a plant pathogen**. *Nature* 1993, **362**:56-58.

11. Roy BA, Bierzychudek P: **The potential for rust infection to cause natural selection in apomictic *Arabis holboellii* (Brassicaceae)**. *Oecologia* 1993, **95**:533-541.

12. Reznick D: **Costs of reproduction: an evaluation of the empirical evidence**. *Oikos* 1985, **44**:257-267.

13. Kraaijeveld AR, Ferrari J, Godfray HCJ: **Costs of resistance in insect-parasite and insect-parasitoid interactions**. *Parasitology* 2002, **125**:S71-S82.

14. Medley GF: **The epidemiological consequences of optimisation of the individual host immune response**. *Parasitology* 2002, **125**:S61-S70.

15. Kaitala V, Heino M, Getz WM: **Host-parasite dynamics and the evolution of host immunity and parasite fecundity strategies**. *B Math Biol* 1997, **59**(3):427-450.

16. Maynard Smith J: **Evolution and the Theory of Games**. Cambridge: Cambridge University Press; 1982.

17. van Baalen M: **Coevolution of recovery ability and virulence**. *Proc R Soc Lond B Biol Sci* 1998, **265**:317-325.

18. Schmid-Hempel P: **Variation in immune defence as a question of evolutionary ecology**. *Proc R Soc Lond B Biol Sci* 2003, **270**:357-366.

19. Lochmiller RL, Deerenberg C: **Trade-offs in evolutionary immunology: just what is the cost of immunity?** *Oikos* 2000, **88**:87-98.

20. Boots M, Begon M: **Trade-offs with resistance to a granulosis virus in Indian meal moth, examined by a laboratory evolution experiment**. *Funct Ecol* 1993, **7**:528-534.

21. Kraaijeveld AR, Godfray HC: **Trade-off between parasitoid resistance and larval competitive ability in *Drosophila melanogaster***. *Nature* 1997, **389**(6648):278-280.

22. Yan G, Severson DW, Chiristensen BM: **Costs and benefits of mosquito refractoriness to malaria parasites: implications for genetic variability of mosquitoes and genetic control of malaria**. *Evolution* 1997, **51**:441-450.

23. Greenman JV, Hudson PJ: **Host exclusion and coexistence in apparent and direct competition: An application of bifurcation theory**. *Theor Popul Biol* 1999, **56**(1):48-64.

24. Schall JJ: **Parasite-mediated competition in Anolis lizards**. *Oecologia* 1992, **92**:58-64.

25. Park T: **Experimental studies of interspecies competition .1. Competition between populations of the flour beetles, *Tribolium confusum* Duval and *Tribolium castaneum* Herbst**. *Ecol Monogr* 1948, **18**(2):265-307.

26. Kiesecker JM, Blaustein AR: **Pathogen reverses competition between larval amphibians**. *Ecology* 1999, **80**(7):2442-2448.

27. Holt RD: **Predation, apparent competition, and structure of prey communities**. *Theor Popul Biol* 1977, **12**(2):197-229.

28. Holt RD, Pickering J: **Infectious-disease and species coexistence - a model of Lotka-Volterra form**. *Am Nat* 1985, **126**(2):196-211.

29. Bonsall MB, Hassell MP: **Apparent competition structures ecological assemblages**. *Nature* 1997, **388**(6640):371-373.

30. Tompkins DM, Greenman JV, Hudson PJ: **Differential impact of a shared nematode parasite on two gamebird hosts: implications for apparent competition**. *Parasitology* 2001, **122**:187-193.

31. Settle WH, Wilson LT: **Invasion by the variegated leafhopper and biotic interactions: parasitism, competition, and apparent competition**. *Ecology* 1990, **71**(4):1461-1470.

32. Torchin ME, Lafferty KD, Kuris AM: **Parasites and marine invasions**. *Parasitology* 2002, **124**:S137-S151.

33. Prenter J, MacNeil C, Dick JTA, Dunn AM: **Roles of parasites in animal invasions**. *Trends Ecol Evol* 2004, **19**(7):385-390.

34. Bedhomme S, Agnew P, Vital Y, Sidobre C, Michalakis Y: **Prevalence-dependent costs of parasite virulence**. *PLoS Biol* 2005, **3**(8):1403-1408.

35. Tompkins DM, White AR, Boots M: **Ecological replacement of native red squirrels by invasive greys driven by disease**. *Ecol Lett* 2003, **6**(3):189-196.

36. Brown SP, Le Chat L, De Paepe M, Taddei F: **Ecology of microbial invasions: Amplification allows virus carriers to invade more rapidly when rare**. *Curr Biol* 2006, **16**(20):2048-2052.

37. Day T, Burns JG: **A consideration of patterns of virulence arising from host-parasite coevolution**. *Evolution* 2003, **57**(3):671-676.

38. Otto SP, Day T: **A Biologist's Guide to Mathematical Modeling**. Princeton, N.J.: Princeton University Press; 2007.

39. Leavy O: **Insect immunity - Immune adaptation in flies**. *Nat Rev Immunol* 2007, **7**(4).

40. Hoffmann JA: **The immune response of *Drosophila***. *Nature* 2003, **426**(6962):33-38.

41. Carton Y, Nappi AJ, Poirie M: **Genetics of anti-parasite resistance in invertebrates**. *Dev Comp Immunol* 2005, **29**(1):9-32.

42. Rolff J, Siva-Jothy MT: **Invertebrate ecological immunology**. *Science* 2003, **301**(5632):472-475.

43. Jokela J, Schmid-Hempel P, Rigby MC: **Dr. Pangloss restrained by the Red Queen - steps towards a unified defence theory**. *Oikos* 2000, **89**(2):267-274.

44. Thrall PH, Burdon JJ: **Host-pathogen dynamics in a metapopulation context: the ecological and evolutionary consequences of being spatial**. *J Ecol* 1997, **85**(6):743-753.

45. Kirchner JW, Roy BA: **The evolutionary advantages of dying young: Epidemiological implications of longevity in metapopulations**. *Am Nat* 1999, **154**(2):140-159.

46. Gandon S, Capowiez Y, Dubois Y, Michalakis Y, Olivieri I: **Local adaptation and gene-for-gene coevolution in a metapopulation model**. *P Roy Soc London B-Biol Sci* 1996, **263**(1373):1003-1009.

47. Thrall PH, Burdon JJ: **Evolution of gene-for-gene systems in metapopulations: the effect of spatial scale of host and pathogen dispersal**. *Plant Pathol* 2002, **51**(2):169-184.

48. Boots M, Sasaki A: **'Small worlds' and the evolution of virulence: infection occurs locally and at a distance**. *Proc R Soc Lond B Biol Sci* 1999, **266**(1432):1933-1938.

49. de Jong G, van Noordwijk AJ: **Acquisition and allocation of resources—genetic (co)variances, selection, and life histories**. *Am Nat* 1992, **139**(4):749-770.

50. Eslin P, Prevost G: **Hemocyte load and immune resistance to *Asobara tabida* are correlated in species of the *Drosophila melanogaster subgroup***. *J Insect Physiol* 1998, **44**(9):807-816.

51. Kraaijeveld AR, Limentani EC, Godfray HC: **Basis of the trade-off between parasitoid resistance and larval competitive ability in *Drosophila melanogaster***. *Proc R Soc Lond B Biol Sci* 2001, **268**(1464):259-261.

52. Little TJ, Kraaijeveld AR: **Ecological and evolutionary implications of immunological priming in invertebrates**. *Trends Ecol Evol* 2004, **19**(2):58-60.

53. Williams PD, Day T: **Interactions between sources of mortality and the evolution of parasite virulence**. *Proc R Soc Lond B Biol Sci* 2001, **268**(1483):2331-2337.

54. Anderson RM, May RM: **Infectious Diseases of Humans**. Oxford: Oxford University Press; 1991.

55. Geritz SAH, Kisdi E, Meszena G, Metz JAJ: **Evolutionarily singular strategies and the adaptive growth and branching of the evolutionary tree**. *Evol Ecol* 1998, **12**(1):35-57.

56. Nee S, May RM: **Dynamics of metapopulations: habitat destruction and competitive coexistence**. *J Anim Ecol* 1992, **61**:37-40.

57. Orlova MV, Smirnova TA, Ganushkina LA, Yacubovich VY, Azizbekyan RR: **Insecticidal activity of *Bacillus laterosporus***. *Appl Environ Microbiol* 1998, **64**(7):2723-2725.

58. Butt T: **Use of entomogenous fungi for the control of insect pests**. In: *Mycota*. Edited by Esser K, Bennett J. Berlin: Springer-Verlag; 2002: 111–134.

59. Anderson RM, May RM: **Regulation and stability of host-parasite population interactions .1. Regulatory processes**. *J Anim Ecol* 1978, **47**(1):219-247.

60. van Boven M, Weissing FJ: **The evolutionary economics of immunity**. *Am Nat* 2004, **163**(2):277-294.

61. Boots M, Bowers RG: **Three mechanisms of host resistance to microparasites - Avoidance, recovery and tolerance - Show different evolutionary dynamics**. *J Theor Biol* 1999, **201**(1):13-23.

62. Bowers RG, Boots M, Begon M: **Life-history trade-offs and the evolution of pathogen resistance: competition between host strains**. *Proc R Soc Lond B* 1994, **257**:247-253.

63. Bowers RG: **The basic depression ratio of the host: the evolution of host resistance to microparasites**. *Proc R Soc Lond B* 2001, **268**:243-250.

64. Boots M, Haraguchi Y: **The evolution of costly resistance in host-parasite systems**. *Am Nat* 1999, **153**:359-370.

65. Graham AL, Allen JE, Read AF: **Evolutionary causes and consequences of immunopathology**. *Annu Rev Ecol Evol Syst* 2005, **36**:373-397.

66. Blair AC, Wolfe LM: **The evolution of an invasive plant: An experimental study with *Silene latifolia***. *Ecology* 2004, **85**(11):3035-3042.

67. Wolfe LM, Elzinga JA, Biere A: **Increased susceptibility to enemies following introduction in the invasive plant *Silene latifolia***. *Ecol Lett* 2004, **7**(9):813-820.

68. Cappuccino N, Carpenter D: **Invasive exotic plants suffer less herbivory than non-invasive exotic plants**. *Biol Lett* 2005, **1**(4):435-438.

69. Wodarz D, Sasaki A: **Apparent competition and recovery from infection**. *J Theor Biol* 2004, **227**:403-412.

70. Hanski IA, Gilpin ME (eds.): **Metapopulation Biology: Ecology, Genetics, and Evolution**. New York: Academic Press; 1997.

71. Hanski I: **Metapopulation Ecology**. Oxford: Oxford University Press; 1999.

72. Haraguchi Y, Sasaki A: **The evolution of parasite virulence and transmission rate in a spatially structured population**. *J Theor Biol* 2000, **203**(2):85-96.

73. Keeling MJ: **The effects of local spatial structure on epidemiological invasions**. *Proc R Soc Lond B Biol Sci* 1999, **266**(1421):859-867.

74. Lively CM, Dybdahl MF: **Parasite adaptation to locally common host genotypes**. *Nature* 2000, **405**(6787):679-681.

75. Rand DA, Keeling M, Wilson HB: **Invasion, stability and evolution to criticality in spatially extended, artificial host-pathogen ecologies**. *P Roy Soc London B-Biol Sci* 1995, **259**(1354):55-63.

76. Pastor-Satorras R, Vespignani A: **Epidemic dynamics and endemic states in complex networks**. *Phys Rev E* 2001, **63**:066117.

77. Pastor-Satorras R, Vespignani A: **Epidemic spreading in scale-free networks**. *Phys Rev Lett* 2001, **86**(14):3200-3203.

78. Pastor-Satorras R, Vespignani A: **Epidemic dynamics in finite size scale-free networks**. *Phys Rev E* 2002, **65**:035108.

79. Zekri N, Clerc JP: **Statistical and dynamical study of disease propagation in a small world network**. *Phys Rev E* 2001, **64**:056115.

80. Jones JH, Handcock MS: **Sexual contacts and epidemic thresholds**. *Nature* 2003, **423**(6940):605-606.

81. Verdasca J, da Gama MMT, Nunes A, Bernardino NR, Pacheco JM, Gomes MC: **Recurrent epidemics in small world networks**. *J Theor Biol* 2005, **233**(4):553-561.

82. Read JM, Keeling MJ: **Disease evolution on networks: the role of contact structure**. *Proc R Soc Lond B Biol Sci* 2003, **270**(1516):699-708.

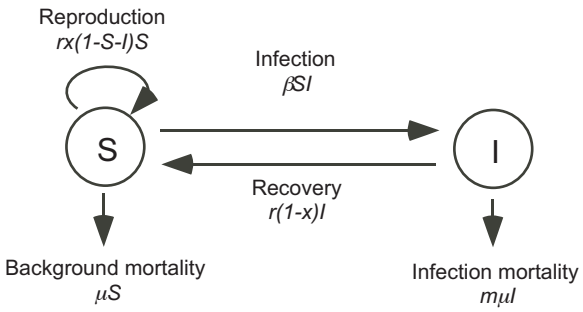


Figure 2.1. Schematic representation of the susceptible-infected-susceptible (*SIS*) model, where r is the resource acquisition rate of the host, x is the proportion of resource allocated to reproduction ($0 < x \leq 1$), β is the parasite transmission rate, μ is the mortality rate of the host, and m is the ratio by which infection increases host mortality.

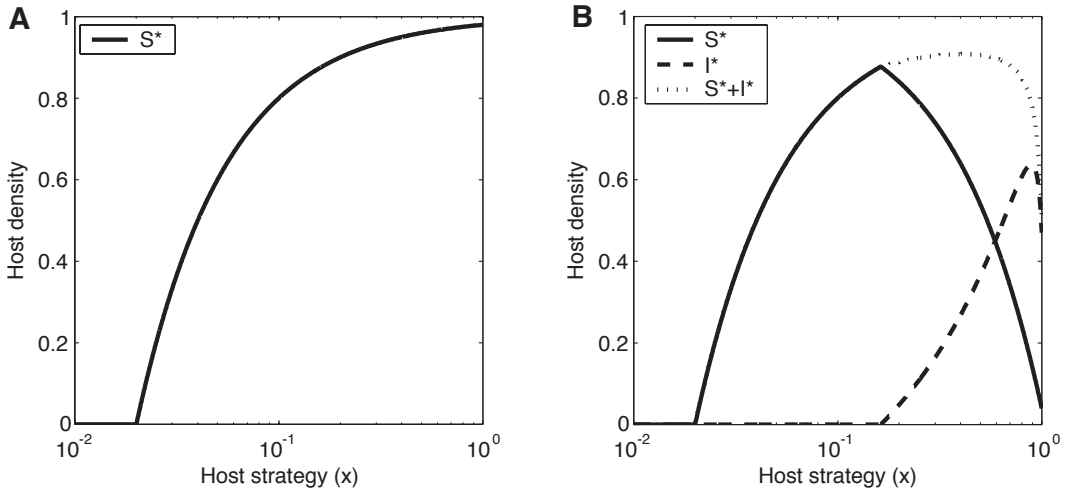


Figure 2.2. Equilibrium host density. (A) In the absence of the parasite, equilibrium host density increases monotonically with investment in reproduction (x). Host population cannot persist if the potential per capita reproduction rate is equal to or lower than the mortality rate ($rx \leq \mu$). (B) When the parasite is present in the environment, the density of susceptible hosts is maximized when the investment in immunity is sufficiently high to eradicate epidemics. Parameter values are: $r = 10$, $\mu = 0.2$, $\beta = 10$, and $m = 2$.

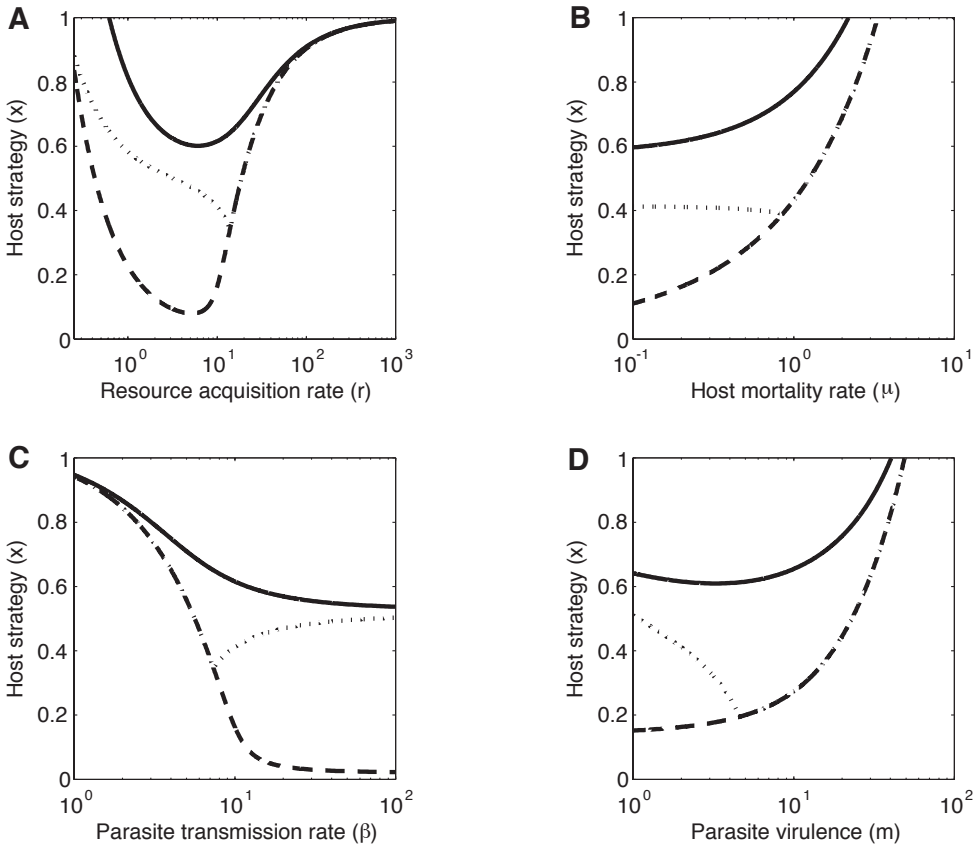


Figure 2.3. Effect of model parameters on host strategies. Dashed lines: the immunizing strategy (x_{imu}); dotted lines: the density maximizing strategy (x_{max}); solid lines: the evolutionarily stable strategy (x_{ESS}). In some part of the parameter space, the immunizing strategy and the density maximizing strategy are equivalent. (A) Host strategies as a function of resource acquisition rate (r). (B) Host strategies as a function of host mortality rate (μ). (C) Host strategies as a function of parasite transmission rate (β). (D) Host strategies as a function of parasite virulence (m). For each comparison, the other parameters are held constant with $r = 10$, $\mu = 0.2$, $\beta = 10$, and $m = 2$.

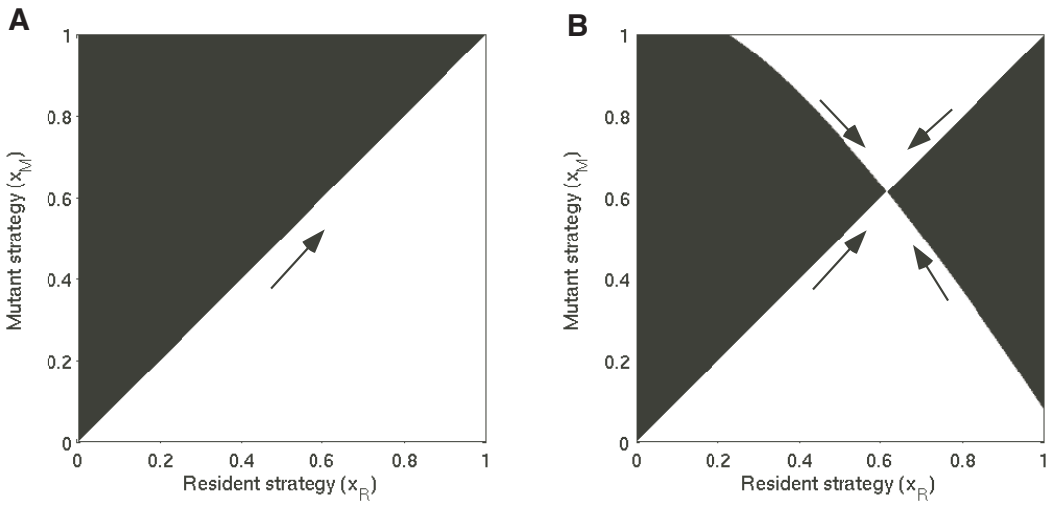


Figure 2.4. Pairwise invasibility plots. The x-axis shows the investment strategy of the resident (x_R) and the y-axis is the strategy of the mutant (x_M). Black regions show the parameter space where the mutant can invade the resident population whereas white regions show the parameter space where the resident is resistant to invasion. Arrows indicate potential evolutionary trajectories of resource allocation strategy in a population. (A) Parasite absent in the population. (B) Parasite present in the population. Parameter values are the same as in Figure 2.2.

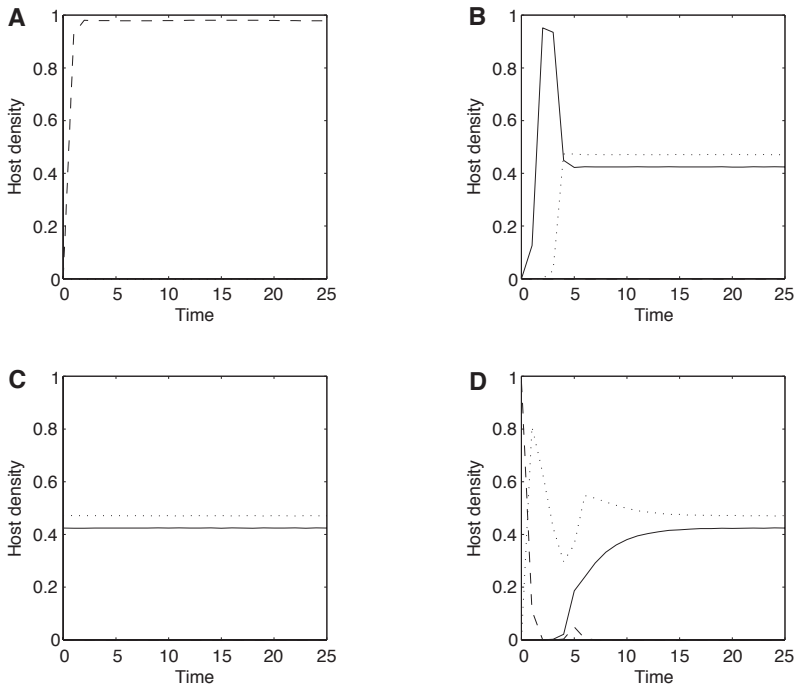


Figure 2.5. Within population dynamics. Dashed lines: density of S_H (susceptible hosts with genotype x_H); solid lines: density of S_P (susceptible hosts with genotype x_P); dotted lines: combined density of I_H and I_P (total density of infected hosts). (A) Colonization of a vacant patch by hosts with genotype x_H. (B) Colonization of a vacant patch by hosts with genotype x_P. (C) Failed invasion of x_H hosts into a patch occupied by x_P hosts. (D) Succesful invasion of x_P hosts into a patch occupied by x_H hosts. For invasion events (panels C and D), the resident hosts occupy the patch at equilibrium density. Hosts with genotype x_H are initiated with a density of S_H = 0.001 and I_H = 0 at time 0 because they are originated from a host-only patch. Hosts with genotype x_P are initiated with a density of S_H = 0 and I_H = 0.001 at time 0 based on the assumption that parasites always accompany x_P hosts in a colonization or invasion event (see text). Parameter values are the same as in Figure 2.2. Under this condition, the generation time of the host is equivalent to five time units.

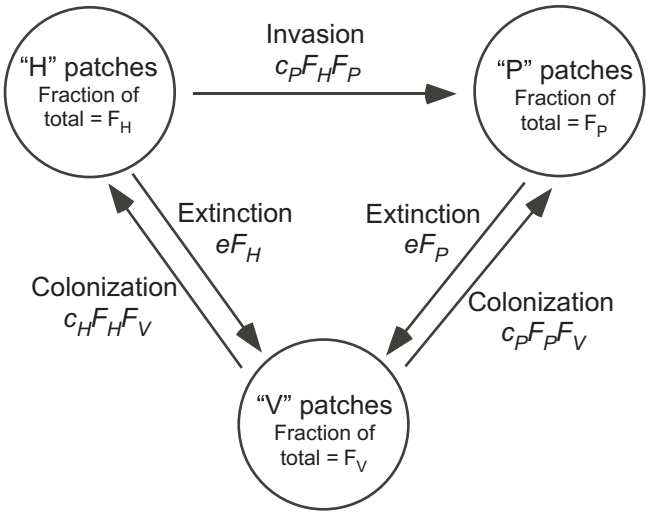


Figure 2.6. Schematic representation of the metapopulation model. The parasite is present in the "P" patches (denoted by a subscript *P*) but not the "H" patches (denoted by a subscript *H* for "host-only"). "V" patches are vacant patches that are available for colonization. c_H and c_P are the colonization rates of the two host genotypes and *e* is the local extinction rate.

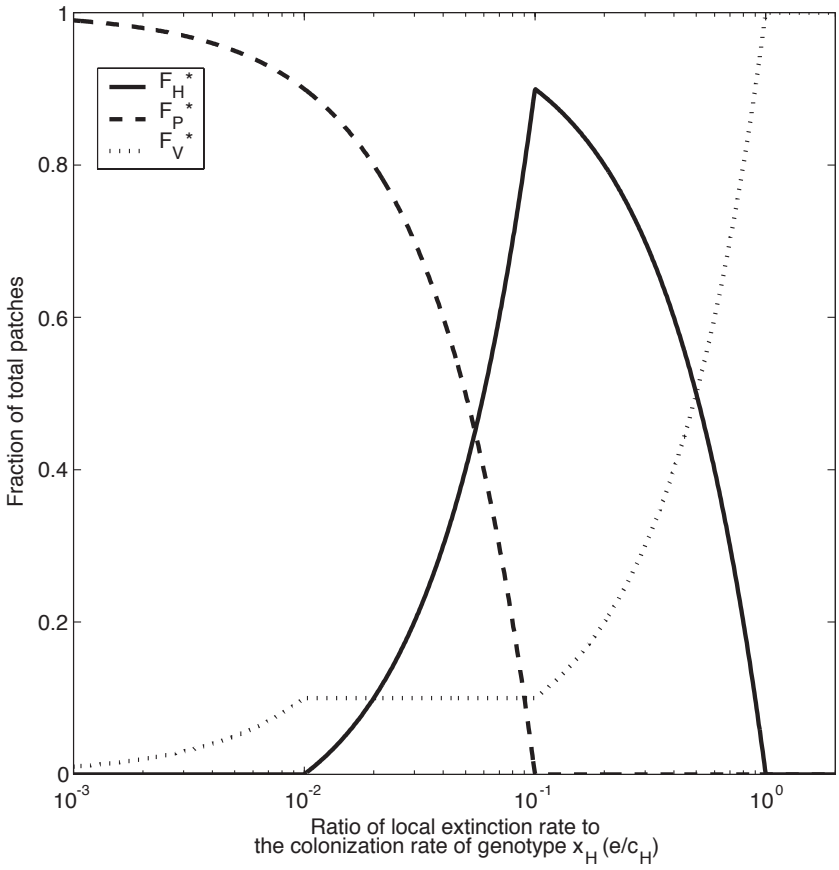


Figure 2.7. Fractions of three patch types at equilibrium in the metapopulation model, as a function of the ratio of local extinction rate to colonization rate of genotype x_H (denoted by e/c_H). Parameter value: $c_P = 0.1c_H$

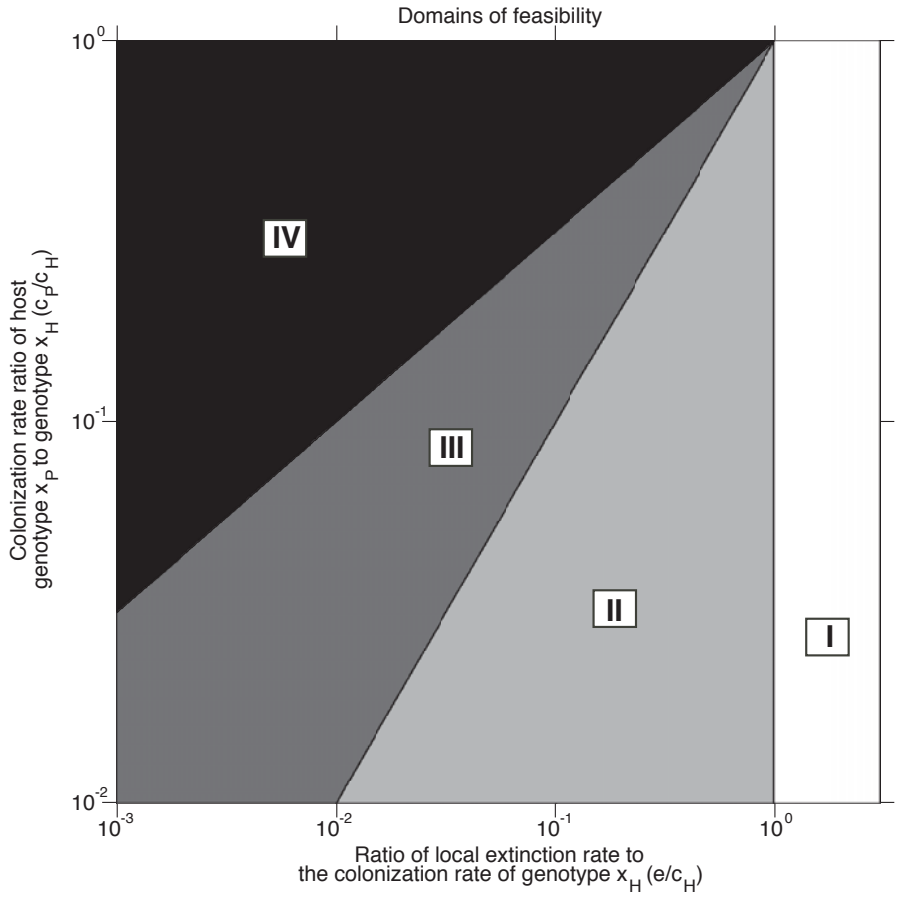


Figure 2.8. Domains of feasibility for the metapopulation model. I. (white region): global extinction ($F_H = 0$ and $F_P = 0$, eq [13]), II. (light gray region): parasite extinction ($F_H > 0$ and $F_P = 0$, eq [14]), III. (dark gray region): parasites present in some local patches ($F_H > 0$ and $F_P > 0$, eq [15]), and IV. (black region): parasites coexist with hosts in all non-vacant patches ($F_H = 0$ and $F_P > 0$, eq [16]).

CHAPTER 3

GENOME-ASSISTED PHYLOGENETICS: IDENTIFICATION OF PHYLOGENETICALLY INFORMATIVE GENES AND THE BIOLOGICAL IMPLICATIONS[1]

[1] Kuo, C.-H., J.P. Wares, and J.C. Kissinger. To be submitted to *Molecular Phylogenetics and Evolution*.

Abstract

Background

Finding the “right sequence” or combination of sequences with the ability to accurately resolve the phylogeny of interest has proven difficult. Often sequence data are painstakingly gathered from many taxa and only after the analysis is completed does it become clear that the sequences did not contain the appropriate resolving power for the question at hand.

Results

We present a new and widely applicable approach to identify phylogenetically informative genes from a set of genome sequences. Our results from two exemplar data sets, Vertebrata and Apicomplexa, demonstrate that the identities of informative genes are highly group-specific, even for phylogenies of similar time scale. We identified many informative genes that are not conventionally employed for phylogenetic inference, including several highly conserved hypothetical proteins. Notably, the level of congruence among gene trees differs greatly between the two taxonomic groups examined; while 89% of the gene trees agree with the species tree in vertebrates, only 36% of the gene trees agree with the species tree in apicomplexans. This result indicates that the minimum number of genes that are necessary to confidently resolve a phylogeny depends on the taxonomic group in question.

Conclusions

Our approach takes advantage of the growing number of available genome sequences to identify a list of genes with the appropriate resolving power for the group of taxa analyzed. The design of our approach to identify informative genes is robust to a number of issues that

normally plague phylogenetic analyses, such as varying GC content, uncertainties of multiple sequence alignment, and violations of assumptions of molecular evolution imposed by phylogenetic algorithms. Identification of phylogenetically informative genes provides useful guidelines for future sequence collection efforts to increase taxon sampling and improve our knowledge about the tree-of-life.

Background

Phylogenetics is the foundation for examining evolutionary processes in any group of organisms. A well-established phylogeny permits inference of the timing of events, such as the origin of traits and the direction of evolutionary change. For most groups of organisms, the available sequence data still limit the resolution of particular taxonomic radiations, and it has become apparent that researchers may in fact be more likely to infer an incorrect phylogenetic tree when there is insufficient data [1, 2]. When the wrong phylogeny is used as the foundation for comparative analyses, the ensuing results will be misleading.

Phylogenetic inference based on molecular sequence data involves several challenges. First, It is often difficult to know which genes one should use *a priori*. Genes differ in their levels of phylogenetic resolving power because they evolve at different rates and under different constraints, therefore the substitution rate of a gene may not be constant across all taxonomic groups. Often sequence data are painstakingly collected from many taxa and only after the analysis is complete does it become apparent that the sequences utilized did not contain appropriate resolving power for the phylogeny in question. Another outcome worse than the generation of uninformative sequence data is the generation of sequence data that are positively misleading and result in strong support for the incorrect phylogenetic tree. Rokas et al. [3] provided a good illustration of these problems in their genome-scale analysis of the yeast phylogeny. Four out of six sequences that are commonly used for phylogenetic inference produced a gene tree with strong bootstrap support for a topology that is different from the species tree. The authors were unable to identify characteristics for predicting the phylogenetic performance of individual genes among the 106 genes examined (but also see [4]).

These challenges arise, in part, because of the potential disparity between gene trees – the reconstruction of ancestral relationship based on individual regions of the genome – and the "true" bifurcating relationship assumed for most organisms. Although gene trees are contained within species trees in the absence of horizontal gene transfer [5], the topology of a gene tree does not necessary match the species tree for reasons such as duplication followed by differential losses [6] and incomplete lineage sorting [7]. With incomplete lineage sorting, genetic polymorphisms in the ancestral species are maintained between speciation events [8]. Unfortunately, there is no clear guideline on which genes, or genomic regions would have an evolutionary history that is reflective of the species phylogeny because of the inherently stochastic nature of incomplete lineage sorting.

In addition to the selection of genes, the technical issues associated with phylogenetic analysis *per se* represent another daunting challenge for inferring phylogeny. It is not uncommon to find that the tree topology supported by a given data set depends on the phylogenetic method used and the associated parameter settings [9-11]. In situations like this, one cannot confidently distinguish which of the alternative topologies reflects the true species phylogeny. Subsequent comparative analysis would then need to consider several competing hypotheses and could result in ambiguous conclusions.

It has been proposed that these problems in molecular phylogenetics can be resolved by using a large number of characters such as those provided by genomic data, assuming there is no systematic bias [12-14]. Several studies have demonstrated that genomic data can resolve a difficult phylogeny with high confidence [12, 15], even when the level of incongruence among gene trees is extremely high because of incomplete lineage sorting [16]. Unfortunately, genomic

data are not available for most organisms of interest and thus have had only a limited impact on the field of molecular phylogenetics to date.

Here we present a high-throughput approach that utilizes existing genomic data to identify phylogenetically informative genes from a group of organisms. We define phylogenetically informative genes as a limited subset of genes that possess a strong phylogenetic signal and are capable of inferring the same tree topology using different phylogenetic methods. Under this definition, phylogenetically informative genes represent the best candidates for sequence collection from additional taxa of interest that lack genome sequence data (Figure 3.1). In contrast to previous genome-scale phylogenetic studies that focused on inferring phylogeny only from organisms with genome sequences [16-18], our approach emphasizes the extraction of information from existing genomic data to facilitate the inference of a more taxonomically comprehensive phylogeny. For this reason, we intentionally discard genes that only contain a weak signal or behave inconsistently under different phylogenetic methods as they are not good candidates for future sequence collection efforts. However, these genes will certainly be of interest under other scenarios.

In this study, we applied our approach to the identification of phylogenetically informative genes to two exemplar eukaryotic data sets, the subphylum Vertebrata and the phylum Apicomplexa (unicellular parasites including the causative agent of malaria, *Plasmodium*). These two data sets are similar in time-scale [19, 20] and we are interested in testing the hypothesis that the sets of informative genes from the two groups would have significant overlap.

In contrast to existing methods that use *a priori* selected genes [21], all available genes [12, 18, 22, 23] or a random subset of the genes [17], our approach provides an objective

procedure to identify genes capable of resolving the phylogeny in question. We begin with the identification of all single-copy orthologous genes from available genomic data for the group of interest. Subsequently, we perform phylogenetic analysis on each orthologous gene set and apply a two-step filtering process to identify genes that contain a high signal-to-noise ratio and have good phylogenetic properties. After these phylogenetically informative genes are identified, we can establish the phylogenetic relationships among the organisms examined based on the consensus of gene trees and/or the concatenated molecular sequences. For this reason, knowledge of the true phylogeny of the selected organisms is not a prerequisite for use of this approach. Furthermore, the amount of molecular sequence data (i.e., the number of genes) that is necessary to confidently resolve a particular phylogeny can be determined on a case-by-case basis depending on the level of congruence among gene trees. Our approach is applicable to taxonomic groups that have at least four representative genome sequences available. With the ever-increasing availability of genome sequences, this approach is applicable to a wide range of taxa.

Results

The vertebrate data set

The vertebrate data set consists of 206,092 protein sequences from seven genomes (Table 3.1), including six vertebrates and one tunicate, *Ciona intestinalis*, as the out-group. We identified 392 single-copy orthologous genes shared by all seven species and from these we obtained 313 alignments that met our usability criteria. OrthoMCL parameter settings had little effect on the number of orthologous genes that satisfied this criterion (data not shown).

We found that 187 genes passed the signal-filtering step and 73 of these also passed the method-filtering step. We consider these 73 genes to be phylogenetically informative because they contain a strong phylogenetic signal and consistently infer the same tree topology (whatever that topology may be) by all three phylogenetic methods used (distance with Neighbor-Joining, maximum parsimony and maximum likelihood). The gene ID, genome location, and functional annotation of these 73 informative genes are listed in Appendix B. When analyzed individually, these 73 informative genes support a total of seven topologies (Figure 3.3). One topology is significantly over represented (65/73, 89%) in comparison to the other 6 topologies that were only represented by one or two genes each (Figure 3.3). The predominant topology represents our current understanding of vertebrate relationships, the species tree [24]. While eight of the informative genes support a gene tree topology that is different from the species tree, none of the informative genes significantly rejects the species tree topology based on the SH test [25] (data not shown). The probability of obtaining a particular gene tree topology based on the given species tree was calculated using COAL [1]. The gene tree topology that is the same as the species tree has only the fifth highest probability among observed gene tree topologies (Figure 3.3).

Phylogenetic analyses to infer the species tree from a consensus of the 73 genes analyzed individually or as a concatenated data alignment produced the same topology (Figure 3.4A). Both of the filtering steps employed increased the level of consensus support at all internal branches. Of note is the observation that all internal branches of the concatenated data alignment received 100% bootstrap support at every stage of the filtering process (Figure 3.4A). The consensus approach and the concatenation approach for inferring the species tree both agree with our prior belief of the “true” species tree.

The apicomplexan data set

The apicomplexan data set consists of 64,974 protein sequences from seven genomes (Table 3.2), including six apicomplexans and one ciliate, *Tetrahymena thermophila*, as the out-group. We identified 441 single-copy orthologous genes and from these we obtained 314 alignments that met our usability criteria. 147 genes passed signal-filtering and 56 of these passed method-filtering (Appendix C). The 56 phylogenetically informative genes support 10 different topologies (Figure 3.5). As was the case with the vertebrate data set, the species tree topology (Figure 3.4B) was best supported in terms of the number of informative genes, but the support is substantially lower (20/56, 35%). In comparison, 7/56 genes support the next best alternative topology. Interestingly, concatenation of the 36 informative genes that do not agree with the species tree topology (10,605 aligned amino acid sites) produces the same topology as the species tree by all three phylogenetic methods (NJ, MP, and MP) with strong bootstrap support (signal strength is 0.97 based on NJ-bootstrap and 0.99 based on MP-bootstrap).

Given the highly reduced genomes of these parasitic organisms, we examined whether or not genes that support the same topology are located in adjacent regions in the genome (i.e., genes support the same topology due to physical linkage). We do not find evidence of spatial clustering, as genes that support the same topology are usually spread across multiple chromosomes (Appendix C). Similar to the vertebrate data set, none of the informative genes significantly reject the species tree topology in the SH test. We found three alternative topologies have a higher probability than the species tree topology when examined with COAL [1] (Figure 3.5).

To test the hypothesis that this high level of incongruence among gene trees was caused in part by a bias in taxon sampling, we performed a taxon removal test to remove one species

from the data set at a time. Figure 3.6 and Table 3.3 summarize the results of all taxon removal tests. Removal (one species at a time) of *Cryptosporidium*, *Theileria*, or *Tetrahymena* greatly reduced the number of observed topologies and increased the proportion of gene trees that agreed with the species tree.

Both the consensus of individual gene trees and analysis of a concatenated data set inferred the same topology for the species tree (Figure 3.4B). Support for the two *Plasmodium* and the two Coccidia (*Eimeria* and *Toxoplasma*) species was strong while support for the two short internal branches was weaker based on the consensus of gene trees. In the analyses of the concatenated alignment, all internal branches received 100% bootstrap support at every stage of the filtering process.

Comparison of the vertebrate and apicomplexan data sets

We found that the numbers of genes at each step of the filtering process were comparable for the two data sets even though the average genome size differed more than three-fold between the data sets (Table 3.4). Despite the similar numbers of informative genes, the two data sets show a sharp contrast in terms of the level of congruence among gene trees. We found that 89% of informative genes supported the species tree in the vertebrate data set whereas only 36% of informative genes supported the species tree in the apicomplexan data set. Interestingly, the informative genes identified in the two data sets both include several highly conserved genes (e.g., DNA replication licensing factors, RNA polymerases, elongation factors, ribosomal proteins) but there is only one overlap, the DNA replication licensing factor MCM5, between the two lists (cf. Appendices B and C). We did not find any identifiable characters that separated the informative genes that agreed with the species tree topology from those that disagreed. In other

words, there is no evidence supporting a systematic process that determined the gene tree topology for a given group of genes. The two groups of genes shared similar amino acid composition in both data sets (data not shown). Furthermore, we observe that several genes that are commonly used for phylogenetic inference disagreed with the species tree in the apicomplexan data set (e.g., actin and a heat shock protein, see Appendix C).

Evaluation of the filtering process

The first filtering step applied to the data was a filter for phylogenetic signal, i.e. the ability of the gene to provide strong support for each node in the tree. Nodal support is assessed by the bootstrap value. In the analyses presented here, we used NJ with bootstrap to estimate the strength of the phylogenetic signal. However, we did assess the effect of Maximum Parsimony (MP) on the inference of these bootstrap values. ML-bootstrap values were not calculated because of the computational cost. The signal strength obtained from the NJ-bootstrap and MP-bootstrap were significantly correlated (R-square = 0.51, P-value = 9.3e-99). In general, MP-bootstrap produced a weaker signal than that of NJ-bootstrap and resulted in a more stringent filtering. The signal strength was significantly correlated with alignment length but the correlation was weak (R-square = 0.11, P-value = 2.5e-18). This result indicates that small genes do not necessarily possess a weaker phylogenetic signal.

CLUSTALW (default settings) without any manual correction was used to generate the alignment before GBLOCKS filtering. To determine the sensitivity to the parameter settings used in sequence alignment, we varied the gap opening penalty by two-fold (i.e., 5, 10 and 20) and compared the resulting NJ trees from all three settings for each gene. Most genes, 73% of the vertebrate and 55% of the apicomplexan, were robust to the gap opening penalty (i.e., inferred

the same NJ tree under all three settings). Only 7% of the vertebrate and 14% of the apicomplexan data sets were sensitive to changes and inferred a different NJ tree under each setting. Most sensitive genes were removed by our filtering process (Table 3.5). Of the 129 informative genes from the two data sets, 112 were robust and only one was sensitive.

In our second filtering step, the test for phylogenetic method robustness, we found that a large proportion of genes, 34% of the vertebrate and 36% of the apicomplexan, were inconsistent with respect to phylogenetic method, inferring a different gene tree topology with each phylogenetic method used (Table 3.6). However, most of these inconsistent genes contain a weak phylogenetic signal and were removed in the signal-filtering step. If we consider only the genes that pass the signal-filtering step, then the genes that infer the same topology by all three phylogenetic methods are the most abundant category for both data sets (Table 3.6).

Each filtering step increased the frequency of genes that support the species tree (Table 3.7). When a gene inferred different tree topologies based on different phylogenetic methods, we found both NJ and MP were more likely to infer the species tree topology in the vertebrate data set while ML performed the best in the apicomplexan data set. During all stages of the filtering process, the believed “true” species tree topology was always the best supported in terms of the frequency of genes. Each filtering step decreased the number of observed gene tree topologies and increased the frequency of genes that supported the species tree topology (Figure 3.7) versus the best alternative topology.

The random gene approach

To test if we could obtain the species tree with high confidence by simply concatenating a large number of genes without any filtering, we performed a random concatenation test. Based on 100 repetitions, a set of 20 randomly chosen genes from all 313 genes in the vertebrate data set produced concatenated alignments that ranged from 4,132 to 6,910 aligned amino acids, with an average of 5,531 sites. All 100 concatenated alignments produced NJ trees that have the same topology as the species tree and with strong bootstrap support. The lowest bootstrap support observed among all branches in these 100 NJ trees was 0.99.

A set of 20 randomly chosen genes from the apicomplexan data set produced concatenated alignments that ranged from 3,806 to 7,447 aligned amino acids, with an average of 5,381 sites. Based on NJ bootstrap, the bootstrap support ranged from 0.78 to 1 with an average of 0.94. However, unlike the vertebrate data set, we found a total of four NJ tree topologies from 100 concatenated alignments. The species tree topology was supported by only 73 of the 100 concatenated alignments.

Discussion

We present an objective and high-throughput approach to determine a set of phylogenetically informative genes when genomic data are available. The phylogenetically informative genes are defined as those genes that have strong bootstrap support and consistently infer the same topology regardless of the phylogenetic methods (NJ, MP, and ML) employed. Our approach of identifying informative genes is generally applicable and flexible. Depending on the phylogeny in question, one can increase or decrease the stringency of selecting informative genes. Once identified, informative genes can be used to determine likely species

tree topologies and as candidate genes for targeted sequencing aimed at increasing taxon sampling.

Based upon analysis of two exemplar data sets of comparable divergence time (vertebrates and apicomplexan protists), we find a striking difference in the level of congruence among gene trees. Moreover, the lists of informative genes from these two groups only share one gene in common despite their similar evolutionary time spans. Our results highlight the fact that there are no universal answers to the number and identity of genes needed to resolve a phylogeny. Nonetheless, our approach of identifying loci that are suitable for inferring species trees provides a solution to these problems. This study presents a highly cost-effective approach that utilizes existing genome data to guide future work in phylogenetics.

The promises and caveats of genome-scale phylogenetics

In phylogenetic inference, we aim to infer the species tree by using gene trees. Although gene trees are contained within species trees in the absence of horizontal gene transfer [5], the topology of a gene tree does not necessarily match the species tree for several reasons: Gene duplication followed by differential losses [6] may lead to misleading analyses from different paralogs; incomplete lineage sorting [7] in which polymorphisms are maintained between speciation events leading to very high levels of incongruence among gene trees under rapid speciation [16]; and GC content has been shown to affect phylogenetic inference both at the nucleotide and protein level [26]. For these reasons, it is crucial to use genes whose evolutionary history is reflective of the species phylogeny. However, it is often not clear what these genes are.

Genomic data offer the prospect that one can reliably recover the true species tree with high confidence as a consequence of the large amount of data, regardless of the genes used for

phylogenetic inference [17]. Several large-scale databases of orthologous genes were created [22, 27], and one even proposed that the process of reconstructing the tree of life using genomic data can be automated [23]. Genomic data can be the key to resolving difficult phylogenies, but there are three caveats: First, available genome data limit us to working with species of poor taxonomic diversity. Second, more data are not necessarily better. Additional data can actually lend stronger support for the wrong tree because of inconsistencies in phylogenetic methods [28-30], multiple substitutions generated by a heterogeneous process [31], nucleotide or amino acid composition bias [13, 31, 32], or poor taxon sampling [33]. Lastly, we often do not know how much data is enough. As illustrated in our results, there is no universal answer to this question. The 20-gene rule suggested by Rokas et al. [3] works well for yeasts, but is probably overkill for vertebrates, and is not enough for the apicomplexans.

The novelties of our approach

Our approach addresses the issues raised above by using existing genomic data (at least four taxonomically appropriate species with genomic or large-scale EST sequence data) to identify phylogenetically informative genes that can be targeted for sequencing in additional taxa lacking a genome sequence. Two recent studies have developed methods that shared our goal of utilizing genomic data to facilitate the phylogenetic inference of under-studied organisms. The first study, by Li *et al.* [34], is focused on the identification of all single-copy exons that are shared by a set of species with genome data available to maximize the number of usable characters. The second method, by Kuramae *et al.* [35], examines the cophenetic correlation coefficients among the distance matrices constructed from single-copy genes for concatenating alignments. Compared to these previous studies, our approach emphasizes the evaluation of

phylogenetic properties of each individual gene and utilizes a series of filtering steps to identify genes suitable for inferring the species tree.

Signal filtering removes genes with a low signal-to-noise ratio because they may evolve at a rate that is too fast or too slow to be suitable for resolving the phylogeny in question. We chose the average bootstrap value across a tree to estimate the strength of phylogenetic signal in this study. However, one can use other criteria such as the minimum or the medium bootstrap value to achieve the same goal. The rationale behind our usage of the average bootstrap value is that it allows genes that contain a strong signal for resolving some clades (but a weak signal for other parts of the tree) to pass the filtering step. Furthermore, the average of estimated clade support was found to be a good indicator of the overall correctness of a tree [36]. We found that bootstrap support estimated via different phylogenetic methods are highly correlated, thus one could utilize the distance method for maximum efficiency in this screening step. When applying this method to different data sets, one can easily increase or decrease the cut-off value to achieve a desirable stringency.

Method filtering screens sequences for violations of underlying assumptions of sequence evolution made by different phylogenetic methods. A method can be inconsistent when the input data violate some of the assumptions, with more data leading toward stronger support for an incorrect tree [37]. This phenomenon presents a dilemma for phylogeneticists because it is difficult to known which tree is truly supported by the data when different methods produce different trees. Each phylogenetic method performs well under different situations. MP is susceptible to long branch attraction while ML is less so [28]. Conversely, ML becomes inconsistent when there exists rate variation within sites (i.e., heterotachy) while MP performs relatively well under a wide range of conditions [29] (but also see [30]). We argue that when all

phylogenetic methods infer the same topology, then it is likely that this topology reflects the true evolutionary history of the sequence because when the phylogenetic signal is strong enough, all methods should produce the same (and presumably correct) tree. In this study we chose the most stringent criterion in method filtering by requiring all three phylogenetic methods to produce the same topology based on a given alignment. We find that a relatively large number of genes pass this filtering step in both data sets. If this filtering criterion is too stringent to produce a sufficient number of loci, one can lower the stringency by allowing a certain degree of disagreement among methods.

Our phylogenetic inference procedure purposefully assumes uniform rates among sites in NJ. The assumption of uniform rates makes NJ very fast for initial screening steps and is not detrimental since we employed a sophisticated Hidden Markov model (HMM) with one invariant plus eight variable rate classes to accommodate rate variation among sites in our ML analyses. We can have high confidence in the gene tree topology when NJ with uniform rate and ML with variable rates both produce the same topology as the result from MP.

We chose OrthoMCL [38] for orthologous gene identification in the first step of our phylogenetic analysis and a benchmark study has confirmed that it performs well [39]. Because OrthoMCL is capable of using incomplete genome data as input and still distinguishes between orthologs and paralogs, we can include species that have a large collection of EST sequences in our analysis and not be restricted to taxa with complete genome sequence. Furthermore, we limited our phylogenetic analysis to single-copy orthologous genes to avoid the problems introduced by paralogs. We performed the orthologous gene identification and sequence alignment at the protein level because of the high level of divergence in nucleotide sequences and bias in GC content present in our taxa. We examined the GC content in the coding region for

the apicomplexan data set and it varied from 24% in *Plasmodium falciparum* to 59% in *Eimeria*. Notably, the two *Plasmodium* species are very different (i.e., *P. falciparium* has a GC content of 24% while *P. vivax* has a GC content of 46%, data not shown). But we always see the two grouped together with strong support. This result indicates that our filtering process is effective and the informative genes perform well.

The correctness of multiple sequence alignment is vital for phylogenetic inference. Manual correction is not suitable for high-throughput analysis because the process is time-consuming, subjective, and not easily reproducible. To avoid problems associated with manual correction while maintaining the quality of alignment, we utilized a programmatic approach to remove regions that contain gaps or are highly divergent. Only highly conserved regions that could be confidently aligned by the alignment programs were used for phylogenetic inference. In addition to being highly scalable and repeatable, this approach effectively reduced the problem of long-branch attraction through the removal of highly divergent regions from alignments. Our results demonstrated that this approach performed well for phylogenetic inference, with more than half of the genes inferring the same tree topology across a wide range of alignment settings. Moreover, most of the genes that were sensitive to alignment settings were eliminated in subsequent steps by our signal and method filtering; only one out of the 129 phylogenetically informative genes was sensitive to alignment settings. Our programmatic approach to alignment using GBLOCKS [40] provides an objective approach to identifying regions of desired conservation.

By examining the level of incongruence among identified informative genes, we can estimate how difficult it will be to resolve the phylogeny in a particular group. In the best-case scenario that the vast majority of informative genes support the species tree (e.g., the vertebrate

data set presented in this study), we would have a high degree of confidence even if we only collect a relatively small number of informative genes from additional taxa to resolve the phylogeny. In contrast, when all observed gene tree topologies are supported by an approximately equal number of genes, we probably cannot confidently resolve the phylogeny even when all informative genes are sequenced from additional taxa. Under this scenario, we may need to seek other approaches to resolve the phylogeny (e.g., gene order [41]) or consider the phylogeny in an alternative form (i.e., not a strict bifurcating tree [42, 43]).

In summary, our methodology provides a high-throughput and objective way to identify a subset of phylogenetically informative genes suitable for resolving a given phylogeny. Variables that affect orthology determination, sequence alignment, gene tree inference, and filtering stringency can all be altered and their effects quantified as needed.

The two exemplar data sets

We have identified phylogenetically informative genes in two distinct taxonomic groups, the vertebrates and apicomplexan protists, each of which have several genome sequences available. The two data sets both cover a time span of approximately 750 million years according to previous studies [19, 20]. Our results confirmed that the two species trees have comparable branch length (Figure 3.4). Surprisingly, the lists of identified informative genes have few overlaps. This result is important because it tells us that there is no universal answer to which genes will be suitable for inferring the species tree even when we know the time-scale that we are dealing with.

Our results indicate that when there are enough data (e.g., genome sequences are available from all taxa), the species tree could be inferred with high confidence. Even for a

difficult phylogeny such as the apicomplexan phylogeny presented in this study, in which some terminal branches were an order of magnitude longer than the internal branches, various approaches based on concatenated alignments and consensus methods all inferred a species tree that is consistent with our prior beliefs [44]. Although it would require further investigation to know whether this is a general rule, this result suggests that knowledge of the true phylogeny of the input genomes is not a prerequisite. Analysis of the phylogenetically informative genes can tell us what the likely species tree is and how difficult it is to resolve. For example, even if we do not know the vertebrate phylogeny *a priori*, the high level of congruence among gene trees tells us that the inferred species tree (in which the consensus and concatenation method produce the same topology) is probably "correct" and the vertebrate phylogeny can be confidently resolved by using fewer genes than that of the apicomplexan phylogeny. Ideally, we would like to test if the phylogenetically informative genes perform well when we increase taxon sampling, but this is not currently feasible due to a lack of data, or lack of data of appropriate taxonomic distance.

None of the informative genes rejected the species tree for either data set in the SH test. These rejection rates are surprisingly low considering the diversity of gene tree topologies observed among all informative genes. One possible explanation for this observation is that these informative genes are single-copy in all lineages and thus are unlikely to involve horizontal gene transfer [45]. Under this hypothesis, these informative genes have been confined in the species tree during their evolutionary history so even when lineage sorting and other processes lead to a gene tree with a different topology than the species tree, the two are unlikely to be radically different. A similar result was obtained by Lerat el al. [18] in their genome-scale phylogenetic analysis of Gamma-proteobacteria. Only two out of 205 single-copy genes rejected

the species tree in their data set and both can be explained by a single horizontal gene transfer event.

For each taxonomic group, we have included an outgroup, *Ciona intestinalis* (a sea squirt) for the vertebrates and *Tetrahymena thermophila* (a ciliate) for the apicomplexans (Tables 3.1 and 3.2). We acknowledge that the choice of *Tetrahymena* as an outgroup for the apicomplexans is not optimal, but there are no available genome data from closer sister taxa such as the dinoflagellates. The existing species tree for the vertebrates is more robust than that of the apicomplexans because of the additional lines of evidence and fossil record [46]. We also know more about the timing of the origin of the genera within vertebrates than among the apicomplexans and thus we were able to choose organisms that were roughly evenly spaced along the species tree. No such data exist for the apicomplexans. We worked with the available data regardless of its distribution along the species tree, and we observed (Figure 3.4B) two extremely short internal branches.

The apicomplexan data set in this study showed a high level of incongruence among gene trees even when we restricted our analysis to genes that passed our stringent filtering process. Only 36% of the genes support the species tree and nine alternative topologies exist. One possible explanation for such a high level of incongruence is that the internal branches are much shorter than the terminal branches in the species tree. As shown by Degnan and Rosenberg [47], a gene tree that has the same topology as the species tree would not have the highest probability among all possible gene trees under this scenario. Among the nine alternative topologies supported by the phylogenetically informative genes in the apicomplexans, we found three alternative topologies that have a higher probability than the species tree, one of which is almost twice as likely (Figure 3.5).

For the vertebrate data set, the species tree has the third lowest probability among the seven observed gene tree topologies. This result could be caused by the highly asymmetrical shape of the species tree [1], as all alternative topologies that have a higher probability are more symmetrical (Figure 3.3). Intriguingly, in both data sets the species tree topology was supported by the highest frequency of genes at all stages of our filtering process despite a relatively low expected probability. The explanation for this discordance between theoretical expectation and experimental observation remains an open question. It is possible that biological processes involve more than neutral character substations, which is one of the assumptions used in the coalescent model. The observation that the topology of species tree is not the most probable under theoretical expectation and yet received the strongest empirical support has two important implications. On the downside, we are likely to obtain an incorrect species tree with strong support when only a limited number of loci were used in inferring trees. Conversely, the true phylogenetic signal will surpass the stochastic noise introduced by lineage sorting when we sample a sufficient number of loci. Both points lead us back to the age-old question in phylogenetic inference: how much data do we need to confidently infer a phylogeny?

How much data do we need?

As simple as it may sound, finding how much data we need to confidently infer a phylogeny is not a trivial task and there might not be a universal answer. Through resampling of 106 genes among eight yeast species, Rokas et al. [3] suggest that 20 randomly selected genes would give the correct species tree with strong support (i.e., a minimum signal strength of 0.95 with a confidence interval of 95%). While we did find a strong signal when we concatenated 20 randomly selected genes in both of our data sets, this random gene approach inferred the correct

phylogeny only 73% of the time and produced three alternative topologies in the apicomplexans. This result serves as a cautionary tale that more data can lead to stronger support for an incorrect answer. It is possible that the random gene approach would perform better with more loci. For example, we obtained the correct species tree with strong support when we concatenated the 36 informative genes that disagreed with the species tree individually. We argue that rather than asking how many randomly selected loci are enough, we should focus on identifying the loci that are appropriate for resolving the phylogenetic problem at hand and collect as much data as possible. For researchers who work on resolving large-scale phylogenies, practical constraints would limit the number of loci one could collect. Our methodology is effective in identifying the genes that have a high signal-to-noise ratio and good phylogenetic properties. By identifying these phylogenetically informative genes prior to data collection, this approach could assist researchers in the allocation of limited resources and increase confidence in the results obtained.

Conclusions

We have shown that different sets of genes are better for resolving particular phylogenies than others even when divergence times for the taxonomic groups in question are similar. Without an informed approach to data collection, we may develop the wrong phylogenetic hypothesis. Given enough data – though it remains unclear how much is "enough" for a given taxon with distinct patterns of species radiation – a robust phylogeny may be developed, but the cost will be prohibitive for some organisms. Since only a small fraction of all described species are currently included in any molecular phylogeny – and typically with only a limited amount of sequence data – it is imperative that we develop more efficient ways to capture the evolutionary diversity present throughout life. Here, we have developed an effective and objective approach

to identify genes that are best able to resolve the phylogeny of a given group, provided that appropriate genomic data from a few representative taxa are available. The conceptual advance behind our methodology, namely an objective strategy for picking the best genes for the job, can provide a highly cost-effective solution to exploring the tree-of-life.

Methods

Data sources and orthologous gene identification

The data sources of the two data sets are listed in Tables 3.1 and 3.2. Our main data analysis procedure is outlined in Figure 3.2. The Perl scripts for performing the analyses described below are freely available from the authors upon request.

Orthologous gene clusters were identified using OrthoMCL [38] (version 1.3, April 10, 2006) with an inflation value (I) of 1.5. 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 [48] to improve sensitivity and specificity. We used WU-BLAST (http://blast.wustl.edu/) (version 2.0) for the all-against-all BLASTP similarity search step. All parameters were defaults except W=5 and E=1e-20 for the vertebrate data set and E=1e-9 for the apicomplexan data set. Following orthologous gene determination by OrthoMCL, only clusters containing single-copy genes were retained for subsequent analysis.

Protein domain identification

Protein domain identification was performed with hmmpfam [49] (version 20.0). *Homo sapiens* was declared as the representative member for each vertebrate orthologous gene and *P. falciparum* for the Apicomplexa since these species have the best available annotation.

Multiple sequence alignment and filtering

CLUSTALW [50] (version 1.83) was used for multiple sequence alignment. We enabled the 'tossgaps' option to ignore gaps when constructing the guide tree. The default gap-opening penalty in CLUSTALW was 10 in both pairwise and multiple alignment steps unless specifically stated otherwise. The alignments produced by CLUSTALW were filtered by GBLOCKS [40] (version 0.91b) to remove regions that contain gaps or are highly divergent. Individual genes that had less than 100 aligned amino acid sites or contained identical sequences from different taxa after GBLOCKS filtering were eliminated from further analysis.

Phylogenetic tree inference

We used the PHYLIP package [51] (version 3.65) for all analyses. The parameters were set to default values unless noted otherwise. NJ trees were constructed using NEIGHBOR with a randomized species input order from a distance matrix created with PROTDIST assuming uniform rates among sites and the Jones, Taylor, and Thornton (JTT) model of amino acid substitutions [52]. MP trees were constructed using PROTPARS with 100 randomizations of input order. When PROTPARS found more than one equally parsimonious tree for a gene, we define the MP tree of that gene to be the strict consensus tree of all equally parsimonious trees.

For ML trees, we first used SiteVarProt [53] (version 1.21) to estimate the substitution rate for each site. The rates were separated into nine bins with one bin for invariant sites and eight equally spaced bins for sites with an estimated rate between zero and one. For each alignment, we added one phantom amino acid site in each bin before calculating the frequency of sites in each bin such that each bin always had a non-zero frequency. This rate variation information was incorporated into ML tree inference by enabling the Hidden Markov Model (HMM) of rate variation option in PROML. In addition, we also enabled the 'global rearrangement' option, performed ten randomizations of input order, and disabled the 'speedier but rougher analysis' option in PROML. The model of amino acid substitution was set to JTT.

For each alignment, we generated 100 bootstrap samples using SEQBOOT. The bootstrap support for each tree was calculated by CONSENSE using extended majority rule. The topology distance between trees was calculated using the symmetric difference [54] implemented in TREEDIST.

Identification of phylogenetically informative genes

We designed a two-step filtering process to identify genes with good phylogenetic properties. The first step of filtering identifies genes with a strong phylogenetic signal. The signal strength of a gene is defined as the average bootstrap support across all internal branches. We used the signal calculated from 100 NJ-bootstrap trees to identify genes that have a signal strength of at least 0.8. To test if the filtering result is sensitive to the phylogenetic method used, we also performed this signal filtering process based on 100 MP-bootstrap trees for each gene.

The second filtering step identifies genes that perform consistently by three phylogenetic methods: distance with Neighbor-Joining (NJ), maximum parsimony (MP), and maximum

likelihood (ML). No assessment of topology is made at this step beyond consistency. For example, gene A may support topology 1 and gene B support topology 2. As long as each gene is consistent in its support of a topology regardless of the phylogenetic method used, it is retained for further analysis. Genes that pass both filtering steps are defined as phylogenetically informative genes.

Inferring the species tree

We use two different approaches to infer the species tree. The first approach is based on the consensus tree produced by analyses of the 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., NJ, MP, and ML) were analyzed separately. The second approach calculated the NJ tree from a concatenated alignment of all genes. For both approaches, we inferred a species tree using three sets of genes: all genes, genes passing signal filtering, and informative genes that passed both filtering steps.

Tree topology test and probability calculation

To ascertain if any of the phylogenetically informative genes significantly rejects the species tree topology, we used the Shimodaira-Hasegawa (SH) test [25] implemented in TREEPUZZLE [55] (version 5.2). The observed gene tree topologies from the phylogenetically informative genes were used as the candidate topologies in the SH test.

For each of the gene tree topologies observed from the phylogenetically informative genes, we calculated the probability of observing the gene tree topology under the given species

tree using COAL [1] (version date: March 27, 2006). The branch length of the species tree was calculated based upon the NJ tree from the concatenated alignment of all informative genes.

Taxon removal analysis

We designed a taxon removal analysis to test the effect of taxon sampling on the level of incongruence among gene trees in the apicomplexan data set. For each test, we removed one taxon from the data set and identified the orthologous genes that were single copy in each of the remaining taxa. Multiple sequence alignment and phylogenetic analysis were as described above. Our taxon removal test is similar to the taxon jackknife method [56] conceptually but was done at the level of orthologous gene selection. In the traditional taxon jackknife method, a taxon is removed after alignment and prior to tree reconstruction. However, the taxon being removed still affects the alignment. We started our taxon removal analysis prior to multiple sequence alignment to eliminate any effects from the taxon being removed.

List of abbreviations

ML, maximum likelihood; MP, maximum parsimony; NJ, neighbor-joining; SH test, Shimodaira-Hasegawa test

Authors' contributions

CHK developed the concept of this study, performed the analysis, and wrote the manuscript. JPW and JCK provided supervision, feedback, suggested additional controls and comments on the manuscript.

Acknowledgements

CHK was supported by a NIH Training Grant (GM07103), the Alton Graduate Research Fellowship, and a Dissertation Completion Assistantship at the University of Georgia. Funding for this work was provided by NIH R01 AI068908 to JCK. P. Brunk, F. Chen, J. Felsenstein, M. Heiges, A. Oliveira, E. Robinson, and H. Wang provided valuable assistance on the use of computer hardware and software. D. Promislow and D. Hall provided helpful comments that improved the manuscript. We thank the Broad Institute, the Dog Genome Sequencing Consortium, DOE Joint Genome Institute, Wellcome Trust Sanger Institute, and the J. Craig Venter Institute for providing pre-publication access to the genome sequence data of *Canis familiaris, Eimeria tenella, Monodelphis domestica, Plasmodium vivax, Toxoplasma gondii,* and *Xenopus tropicalis.* Sequencing of *Toxoplasma gondii* was funded by the National Institute of Allergy and Infectious Disease.

References

1. Degnan JH, Salter LA: **Gene tree distributions under the coalescent process**. *Evolution* 2005, **59**(1):24-37.

2. Degnan JH, Rosenberg NA: **Discordance of species trees with their most likely gene trees**. *PLoS Genet* 2006, **2**(5):762-768.

3. Holder M, Lewis PO: **Phylogeny estimation: Traditional and Bayesian approaches**. *Nat Rev Genet* 2003, **4**(4):275-284.

4. Collins TM, Fedrigo O, Naylor GJP: **Choosing the best genes for the job: The case for stationary genes in genome-scale phylogenetics**. *Syst Biol* 2005, **54**(3):493-500.

5. Gogarten JP, Townsend JP: **Horizontal gene transfer, genome innovation and evolution**. *Nat Rev Microbiol* 2005, **3**(9):679-687.

6. Baldauf SL: **Phylogeny for the faint of heart: a tutorial**. *Trends Genet* 2003, **19**(6):345-351.

7. Maddison WP: **Gene trees in species trees**. *Syst Biol* 1997, **46**(3):523-536.

8. Pamilo P, Nei M: **Relationships between gene trees and species trees**. *Mol Biol Evol* 1988, **5**(5):568-583.

9. Russo CAM, Takezaki N, Nei M: **Efficiencies of different genes and different tree-building methods in recovering a known vertebrate phylogeny**. *Mol Biol Evol* 1996, **13**(3):525-536.

10. Swofford DL, Waddell PJ, Huelsenbeck JP, Foster PG, Lewis PO, Rogers JS: **Bias in phylogenetic estimation and its relevance to the choice between parsimony and likelihood methods**. *Syst Biol* 2001, **50**(4):525-539.

11. Huelsenbeck JP: **Performance of phylogenetic methods in simulation**. *Syst Biol* 1995, **44**(1):17-48.

12. Philippe H, Lartillot N, Brinkmann H: **Multigene analyses of bilaterian animals corroborate the monophyly of Ecdysozoa, Lophotrochozoa, and Protostomia**. *Mol Biol Evol* 2005, **22**(5):1246-1253.

13. Phillips MJ, Delsuc FD, Penny D: **Genome-scale phylogeny and the detection of systematic biases**. *Mol Biol Evol* 2004, **21**(7):1455-1458.

14. Rokas A: **Genomics and the tree of life**. *Science* 2006, **313**(5795):1897-1899.

15. Philip GK, Creevey CJ, McInerney JO: **The Opisthokonta and the Ecdysozoa may not be clades: Stronger support for the grouping of plant and animal than for animal and fungi and stronger support for the Coelomata than Ecdysozoa**. *Mol Biol Evol* 2005, **22**(5):1175-1184.

16. Pollard DA, Iyer VN, Moses AM, Eisen MB: **Widespread discordance of gene trees with species tree in *Drosophila*: Evidence for incomplete lineage sorting**. *PLoS Genet* 2006, **2**(10):1634-1647.

17. Rokas A, Williams BL, King N, Carroll SB: **Genome-scale approaches to resolving incongruence in molecular phylogenies**. *Nature* 2003, **425**(6960):798-804.

18. Lerat E, Daubin V, Moran NA: **From gene trees to organismal phylogeny in prokaryotes: the case of the gamma-proteobacteria**. *PLoS Biol* 2003, **1**(1):101-109.

19. Douzery EJP, Snell EA, Bapteste E, Delsuc F, Philippe H: **The timing of eukaryotic evolution: Does a relaxed molecular clock reconcile proteins and fossils?** *Proc Natl Acad Sci USA* 2004, **101**(43):15386-15391.

20. Blair JE, Shah P, Hedges SB: **Evolutionary sequence analysis of complete eukaryote genomes**. *BMC Bioinformatics* 2005, **6**.

21. Hillis D, Moritz C, Mable B (eds.): **Molecular Systematics**, 2nd edn. Sunderland, Massachusetts: Sinauer Associates; 1996.

22. Dehal PS, Boore JL: **A phylogenomic gene cluster resource: the Phylogenetically Inferred Groups (PhIGs) database**. *BMC Bioinformatics* 2006, **7**.

23. Ciccarelli FD, Doerks T, von Mering C, Creevey CJ, Snel B, Bork P: **Toward automatic reconstruction of a highly resolved tree of life**. *Science* 2006, **311**(5765):1283-1287.

24. Stuart GW, Moffett K, Leader JJ: **A comprehensive vertebrate phylogeny using vector representations of protein sequences from whole genomes**. *Mol Biol Evol* 2002, **19**(4):554-562.

25. Shimodaira H, Hasegawa M: **Multiple comparisons of log-likelihoods with applications to phylogenetic inference**. *Mol Biol Evol* 1999, **16**(8):1114-1116.

26. Foster PG, Hickey DA: **Compositional bias may affect both DNA-based and protein-based phylogenetic reconstructions**. *J Mol Evol* 1999, **48**(3):284-290.

27. Chen F, Mackey AJ, Stoeckert CJ, Roos DS: **OrthoMCL-DB: querying a comprehensive multi-species collection of ortholog groups**. *Nucleic Acids Res* 2006, **34**:D363-D368.

28. Huelsenbeck JP: **Systematic bias in phylogenetic analysis: Is the Strepsiptera problem solved?** *Syst Biol* 1998, **47**(3):519-537.

29. Kolaczkowski B, Thornton JW: **Performance of maximum parsimony and likelihood phylogenetics when evolution is heterogeneous**. *Nature* 2004, **431**(7011):980-984.

30. Philippe H, Zhou Y, Brinkmann H, Rodrigue N, Delsuc F: **Heterotachy and long-branch attraction in phylogenetics**. *BMC Evol Biol* 2005, **5**.

31. Jeffroy O, Brinkmann H, Delsuc F, Philippe H: **Phylogenomics: the beginning of incongruence?** *Trends Genet* 2006, **22**(4):225-231.

32. Lockhart PJ, Steel MA, Hendy MD, Penny D: **Recovering evolutionary trees under a more realistic model of sequence evolution**. *Mol Biol Evol* 1994, **11**(4):605-612.

33. Soltis DE, Albert VA, Savolainen V, Hilu K, Qiu YL, Chase MW, Farris JS, Stefanovic S, Rice DW, Palmer JD *et al*: **Genome-scale data, angiosperm relationships, and 'ending incongruence': a cautionary tale in phylogenetics**. *Trends Plant Sci* 2004, **9**(10):477-483.

34. Li C, Orti G, Zhang G, Lu G: **A practical approach to phylogenomics: the phylogeny of ray-finned fish (Actinopterygii) as a case study**. *BMC Evol Biol* 2007, **7**(1):44.

35. Kuramae E, Robert V, Echavarri-Erasun C, Boekhout T: **Cophenetic correlation analysis as a strategy to select phylogenetically informative proteins: an example from the fungal kingdom**. *BMC Evol Biol* 2007, **7**(1):134.

36. Hall BG, Salipante SJ: **Measures of clade confidence do not correlate with accuracy of phylogenetic trees**. *PLoS Comput Biol* 2007, **3**(3):e51.

37. Felsenstein J: **Cases in which parsimony or compatibility methods will be positively misleading**. *Syst Zool* 1978, **27**(4):401-410.

38. Li L, Stoeckert CJ, Roos DS: **OrthoMCL: Identification of ortholog groups for eukaryotic genomes**. *Genome Res* 2003, **13**(9):2178-2189.

39. Hulsen T, Huynen MA, de Vlieg J, Groenen PMA: **Benchmarking ortholog identification methods using functional genomics data**. *Genome Biol* 2006, **7**(4).

40. Castresana J: **Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis**. *Mol Biol Evol* 2000, **17**(4):540-552.

41. Snel B, Huynen MA, Dutilh BE: **Genome trees and the nature of genome evolution**. *Annu Rev Microbiol* 2005, **59**:191-209.

42. Kunin V, Goldovsky L, Darzentas N, Ouzounis CA: **The net of life: Reconstructing the microbial phylogenetic network**. *Genome Res* 2005, **15**(7):954-959.

43. Dagan T, Martin W: **The tree of one percent**. *Genome Biol* 2006, **7**(10):118.

44. Morrison DA, Ellis JT: **Effects of nucleotide sequence alignment on phylogeny estimation: A case study of 18S rDNAs of Apicomplexa**. *Mol Biol Evol* 1997, **14**(4):428-441.

45. Daubin V, Gouy M, Perriere G: **A phylogenomic approach to bacterial phylogeny: Evidence of a core of genes sharing a common history**. *Genome Res* 2002, **12**(7):1080-1090.

46. Kumar S, Hedges SB: **A molecular timescale for vertebrate evolution**. *Nature* 1998, **392**(6679):917-920.

47. Embley TM, Martin W: **Eukaryotic evolution, changes and challenges**. *Nature* 2006, **440**(7084):623-630.

48. Van Dongen S: **Graph clustering by flow simulation**. *PhD thesis.* University of Utrecht; 2000.

49. Bateman A, Coin L, Durbin R, Finn RD, Hollich V, Griffiths-Jones S, Khanna A, Marshall M, Moxon S, Sonnhammer ELL *et al*: **The Pfam protein families database**. *Nucleic Acids Res* 2004, **32**:D138-D141.

50. Thompson JD, Higgins DG, Gibson TJ: **CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, positions-specific gap penalties and weight matrix choice**. *Nucleic Acids Res* 1994, **22**:4673 - 4680.

51. Felsenstein J: **PHYLIP (Phylogeny Inference Package)** In., Version 3.65 edn. Seattle: Department of Genome Sciences, University of Washington; 2005.

52. Jones DT, Taylor WR, Thornton JM: **The rapid generation of mutation data matrices from protein sequences**. *Comput Appl Biosci* 1992, **8**(3):275-282.

53. Horner DS, Pesole G: **The estimation of relative site variability among aligned homologous protein sequences**. *Bioinformatics* 2003, **19**(5):600-606.

54. Robinson DF, Foulds LR: **Comparison of phylogenetic trees**. *Math Biosci* 1981, **53**(1-2):131-147.

55. Schmidt HA, Strimmer K, Vingron M, von Haeseler A: **TREE-PUZZLE: maximum likelihood phylogenetic analysis using quartets and parallel computing**. *Bioinformatics* 2002, **18**(3):502-504.

56. Siddall ME: **Another monophyly index: Revisiting the jackknife**. *Cladistics* 1995, **11**(1):33-56.

57. Hubbard TJP, Aken BL, Beal K, Ballester B, Caccamo M, Chen Y, Clarke L, Coates G, Cunningham F, Cutts T *et al*: **Ensembl 2007**. *Nucleic Acids Res* 2007, **35**(suppl_1):D610-617.

58. Heiges M, Wang HM, Robinson E, Aurrecoechea C, Gao X, Kaluskar N, Rhodes P, Wang S, He CZ, Su YQ *et al*: **CryptoDB: a *Cryptosporidium* bioinformatics resource update**. *Nucleic Acids Research* 2006, **34**:D419-D422.

59. Hertz-Fowler C, Peacock CS, Wood V, Aslett M, Kerhornou A, Mooney P, Tivey A, Berriman M, Hall N, Rutherford K *et al*: **GeneDB: a resource for prokaryotic and eukaryotic organisms**. *Nucleic Acids Res* 2004, **32**(suppl_1):D339-343.

60. Bahl A, Brunk B, Crabtree J, Fraunholz MJ, Gajria B, Grant GR, Ginsburg H, Gupta D, Kissinger JC, Labo P *et al*: **PlasmoDB: the *Plasmodium* genome resource. A database integrating experimental and computational data**. *Nucleic Acids Res* 2003, **31**(1):212-215.

61. Gajria B, Bahl A, Brestelli J, Dommer J, Fischer S, Gao X, Heiges M, Iodice J, Kissinger JC, Mackey AJ *et al*: **ToxoDB: an integrated *Toxoplasma gondii* database resource**. *Nucleic Acids Res* 2007:gkm981.

62. **J. Craig Venter Institute** [http://jcvi.org/]

Table 3.1. The vertebrate data set. The sea *squirt Ciona intestinalis* is included as the outgroup.

Abbr.	Species name	Number of sequences	Version date	Data source
Cf	*Canis familiaris*	30,321	07/27/2006	Ensembl [57]
Gg	*Gallus gallus*	24,168	07/31/2006	Ensembl [57]
Hs	*Homo sapiens*	48,926	08/01/2006	Ensembl [57]
Md	*Monodelphis domestica*	32,251	08/02/2006	Ensembl [57]
Tr	*Takifugu rubripes*	22,102	08/02/2006	Ensembl [57]
Xt	*Xenopus tropicalis*	28,324	08/02/2006	Ensembl [57]
Ci	*Ciona intestinalis*	20,000	07/27/2006	Ensembl [57]

Table 3.2. The apicomplexan data set. The ciliate *Tetrahymena thermophila* is included as the outgroup.

Abbr.	Species name	Number of sequences	Version date	Data source
Cp	*Cryptosporidium parvum*	3,806	04/02/2006	CryptoDB [58]
Et	*Eimeria tenella*	11,393	05/22/2006	GeneDB [59]
Pf	*Plasmodium falciparum*	5,411	12/07/2005	PlasmoDB [60]
Pv	*Plasmodium vivax*	5,352	12/07/2005	PlasmoDB [60]
Ta	*Theileria annulata*	3,795	07/15/2005	GeneDB [59]
Tg	*Toxoplasma gondii*	7,793	01/04/2006	ToxoDB [61]
Tt	*Tetrahymena thermophila*	27,424	04/14/2006	J. Craig Venter Institute [62]

Table 3.3. Effect of taxon removal in the apicomplexan data set. The column 'ALL' refers to the original 7-taxon data set and each taxon removal data set is denoted by a '-' sign before the species name abbreviation.

Data set	ALL	*-Cp*	*-Et*	*-Pf*	*-Pv*	*-Ta*	*-Tg*	*-Tt*
Number of single-copy orthologous genes	441	492	715	443	447	475	446	563
Number of alignments	314	326	563	284	288	307	287	368
Number of genes after signal filtering	147	261	202	71	78	226	111	289
Number of informative genes	56	141	112	37	34	128	50	183
Number of gene tree topologies	10	6	13	9	10	4	12	5
Frequency of genes supporting the species tree topology	0.36	0.44	0.28	0.38	0.44	0.54	0.34	0.65
Frequency of genes supporting the best alternative topology	0.13	0.31	0.17	0.19	0.15	0.31	0.14	0.19

Table 3.4. Summary of results using all taxa.

Data set	Vertebrate	Apicomplexan
Number of protein sequences	206,092	64,974
Number of single-copy orthologous genes	392	441
Number of alignments	313	314
Number of genes after signal filtering	187	147
Number of informative genes	73	56
Number of gene tree topologies	7	10
Frequency of genes supporting the species tree topology	0.89	0.36
Frequency of genes supporting the best alternative topology	0.03	0.13

Table 3.5. Sensitivity to multiple sequence alignment settings. A gene is defined as robust when it inferred the same NJ tree topology under all three gap opening settings, whereas a gene is defined as sensitive when it inferred a different NJ tree topology under each of the gap opening settings. The genes in the intermediate class inferred the same NJ tree topology under two different gap opening settings but a third setting produced a different NJ tree topology. The numbers show the count of genes in each class, with percentages in parentheses.

Class	Vertebrate			Apicomplexan		
	All (313 genes)	Post signal filtering (187 genes)	Informative (73 genes)	All (314 genes)	Post signal filtering (147 genes)	Informative (56 genes)
Robust	230 (73%)	163 (87%)	66 (90%)	173 (55%)	109 (74%)	46 (82%)
Intermediate	62 (20%)	21 (11%)	7 (10%)	98 (31%)	27 (18%)	9 (16%)
Sensitive	21 (7%)	3 (2%)	0 (0%)	43 (14%)	11 (7%)	1 (2%)

Table 3.6. Agreement among phylogenetic methods. The numbers indicate counts of genes in each category, with percentages in parentheses. The most abundant category in each data set is highlighted in bold. The inferred topology can be any topology so long as it is consistent within the methods being evaluated.

Category	Vertebrate		Apicomplexan	
	All (313 genes)	Post signal filtering (187 genes)	All (314 genes)	Post signal filtering (147 genes)
All methods inferred the same topology	82 (26%)	**73 (39%)**	70 (22%)	**56 (38%)**
NJ and MP inferred the same topology	71 (23%)	48 (26%)	39 (12%)	24 (16%)
NJ and ML inferred the same topology	35 (11%)	15 (8%)	66 (21%)	26 (18%)
MP and ML inferred the same topology	20 (6%)	14 (7%)	25 (8%)	11 (7%)
Each method inferred a different topology	**105 (34%)**	37 (20%)	**114 (36%)**	30 (20%)

Table 3.7. Performance of phylogenetic methods. The numbers show the counts of genes that infer the species tree, with percentages in parentheses.

Method	Vertebrate			Apicomplexan		
	All (313 genes)	Post signal filtering (187 genes)	Informative (73 genes)	All (314 genes)	Post signal filtering (147 genes)	Informative (56 genes)
NJ	164 (52%)	126 (67%)	65 (89%)	51 (16%)	33 (22%)	20 (36%)
MP	161 (51%)	129 (69%)	65 (89%)	44 (14%)	30 (20%)	20 (36%)
ML	103 (33%)	87 (47%)	65 (89%)	61 (19%)	45 (31%)	20 (36%)

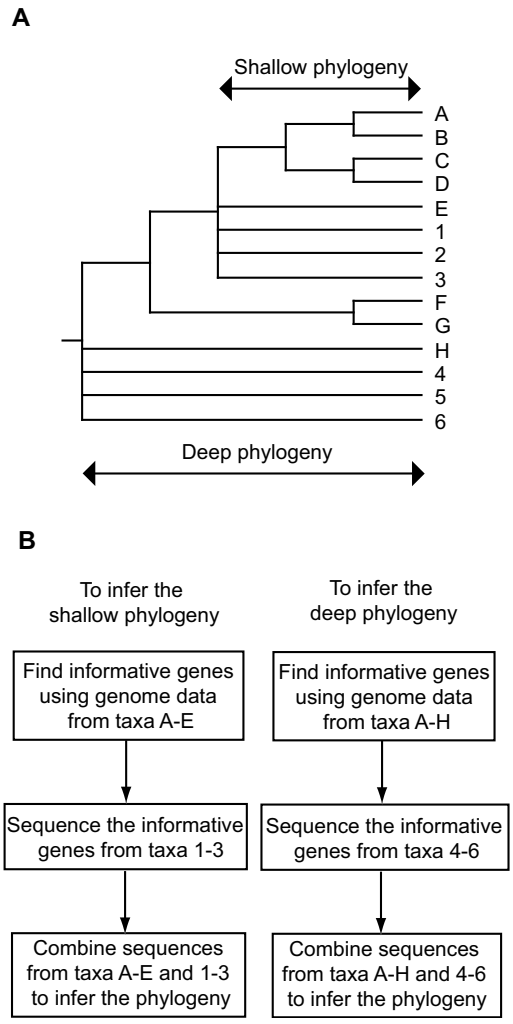


Figure 3.1. Methodological concept. A. A hypothetical phylogeny. Taxa A-H are the species with available genomic sequence data. The phylogenetic placement of taxa 1-6 are yet to be determined and no genomic sequence is available. B. The methodological flow chart. To obtain a shallow phylogeny, genomic data from taxa A-E can be used as the input for our methodology to infer a species tree and identify the phylogenetically informative genes. After the phylogenetically informative genes are identified, the orthologous sequences from taxa 1-3 can be obtained and used to infer the complete phylogeny. Similarly, genome data from taxa A-H can be used as the input when one is interested in the 'deep phylogeny' and the phylogenetic placement of taxa 4-6.

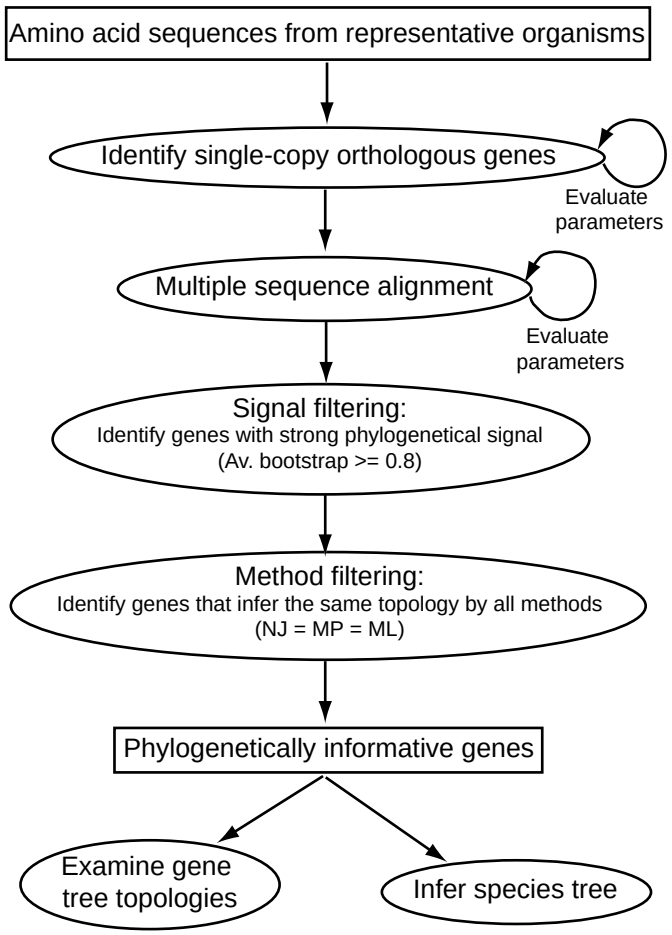


Figure 3.2. Flow chart of the analysis procedure. Please refer to the materials and methods for exact methods and parameters evaluated.

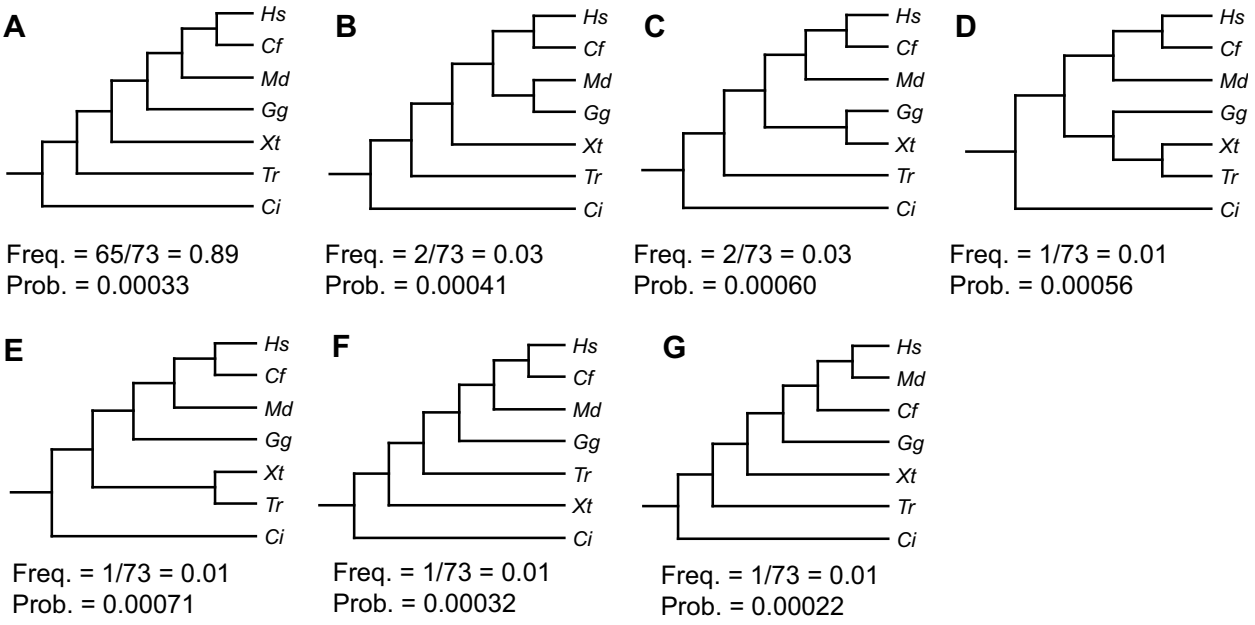


Figure 3.3. Vertebrate gene tree topologies. Numbers below each panel indicate the frequency of genes that support the indicated topology (Freq.) and the probability of obtaining the gene tree topology given the species tree (Prob.). *Hs = Homo sapiens; Cf =Canis familiaris; Md = Monodelphis domestica; Gg = Gallus gallus; Xt =Xexopus tropicalis; Tr =Takifugo rubripes; Ci = Ciona intestinalis.*

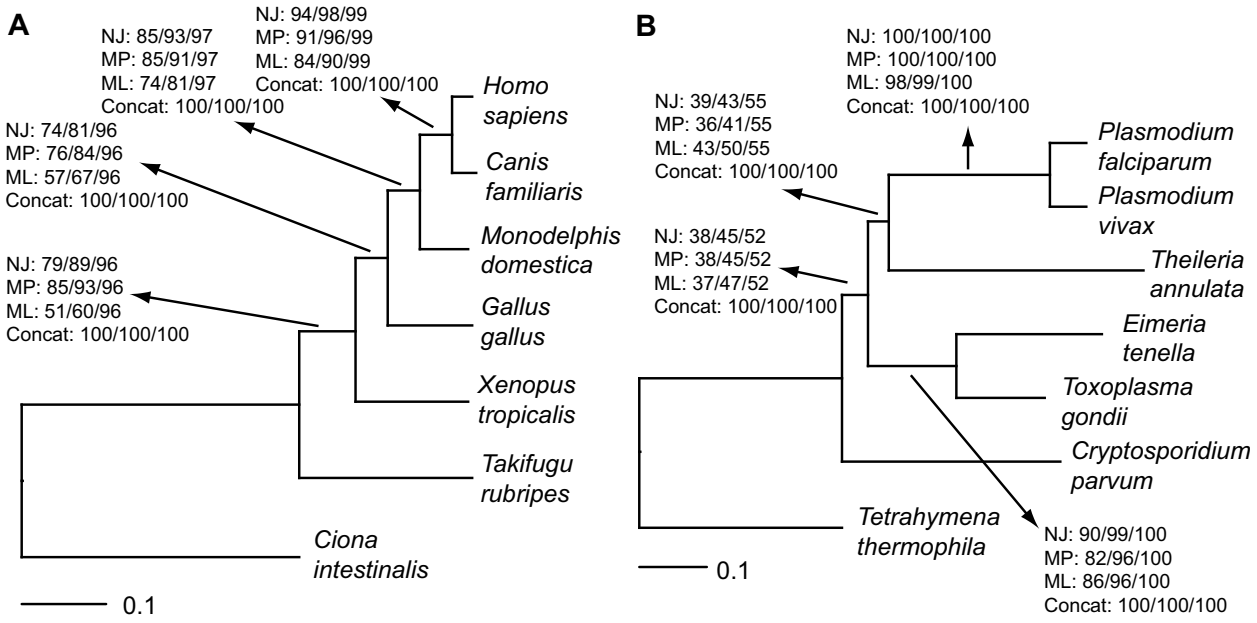


Figure 3.4. The species trees. Branch lengths are based on the NJ tree generated from the concatenated alignment of all informative genes. Internal branch labels indicate the level of bootstrap support inferred by different phylogenetic methods. First row: consensus support based on NJ gene trees; second row: MP gene trees; third row: ML gene trees; fourth row: bootstrap support from a concatenated alignment analyzed with NJ method. Within each row, the sets of numbers between slashes are the level of support calculated from (1) all usable genes, (2) genes passing the signal-filtering step, and (3) identified phylogenetically informative genes (passed both filtering steps). A. The species tree for the vertebrate data set. B. The species tree for the apicomplexan data set.

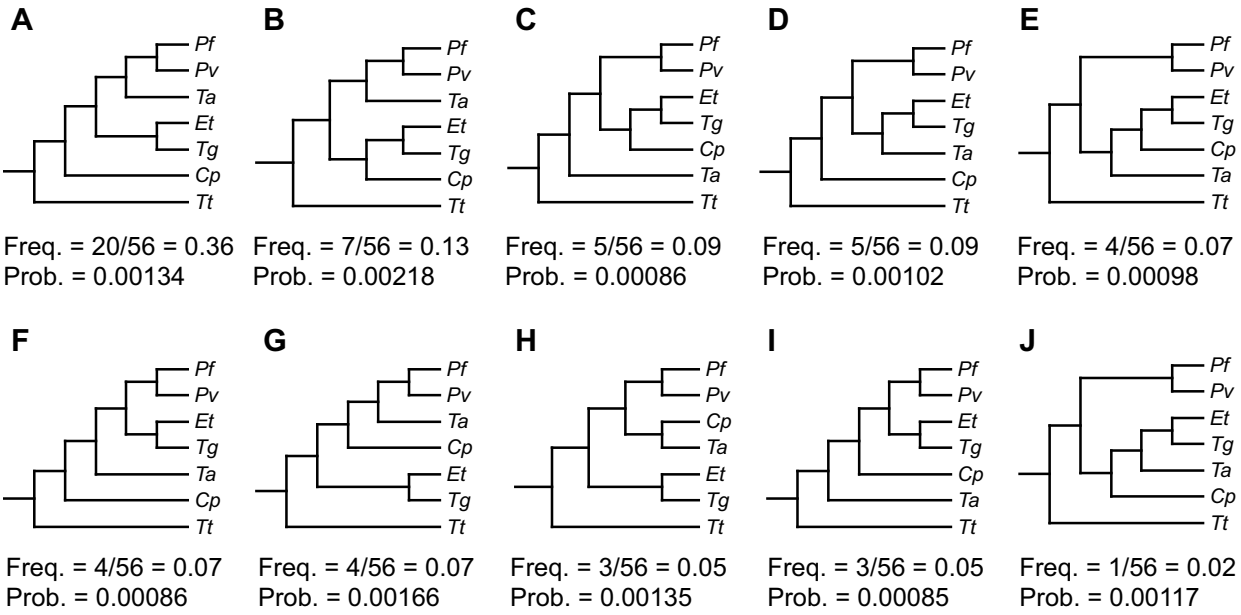


Figure 3.5. Apicomplexan gene tree topologies. Numbers below each panel indicate the frequency of genes that support the topology (Freq.) and the probability of obtaining the gene tree topology given the species tree (Prob.). *Pf = Plasmodium falciparum; Pv = Plasmodium vivax; Ta = Theileria annulata; Et = Eimeria tenella; Tg = Toxoplasma gondii; Cp = Cryptosporidium parvum; Tt = Tetrahymena thermophila.*

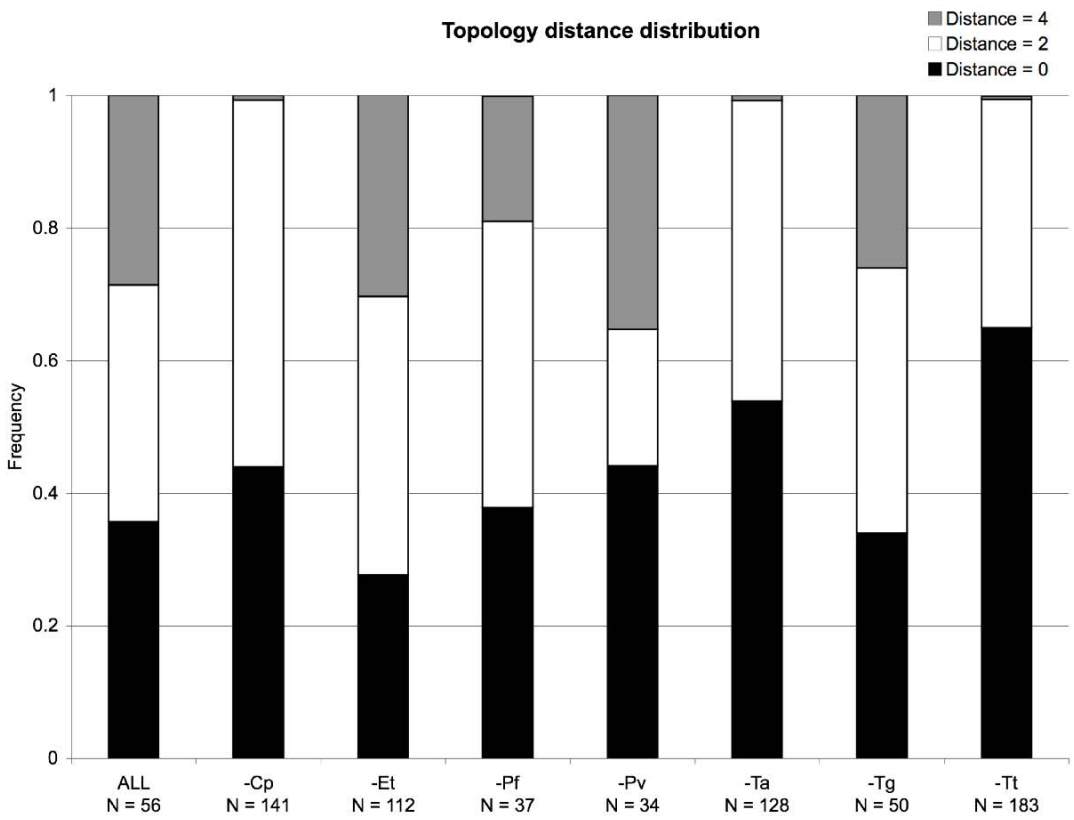


Figure 3.6. Effect of taxon removal on topology distance distribution. 'ALL' refers to the original 7-taxon data set and each taxon removal data set is denoted by a '-' sign before the species name abbreviation for the taxon that was removed. The number of informative genes found in each data set is labeled on the X-axis. The topological distance between the gene trees and the species tree is measured by symmetrical differences as implemented in TREEDIST. The black bar indicates the frequency of genes that infer the same topology as the species tree.

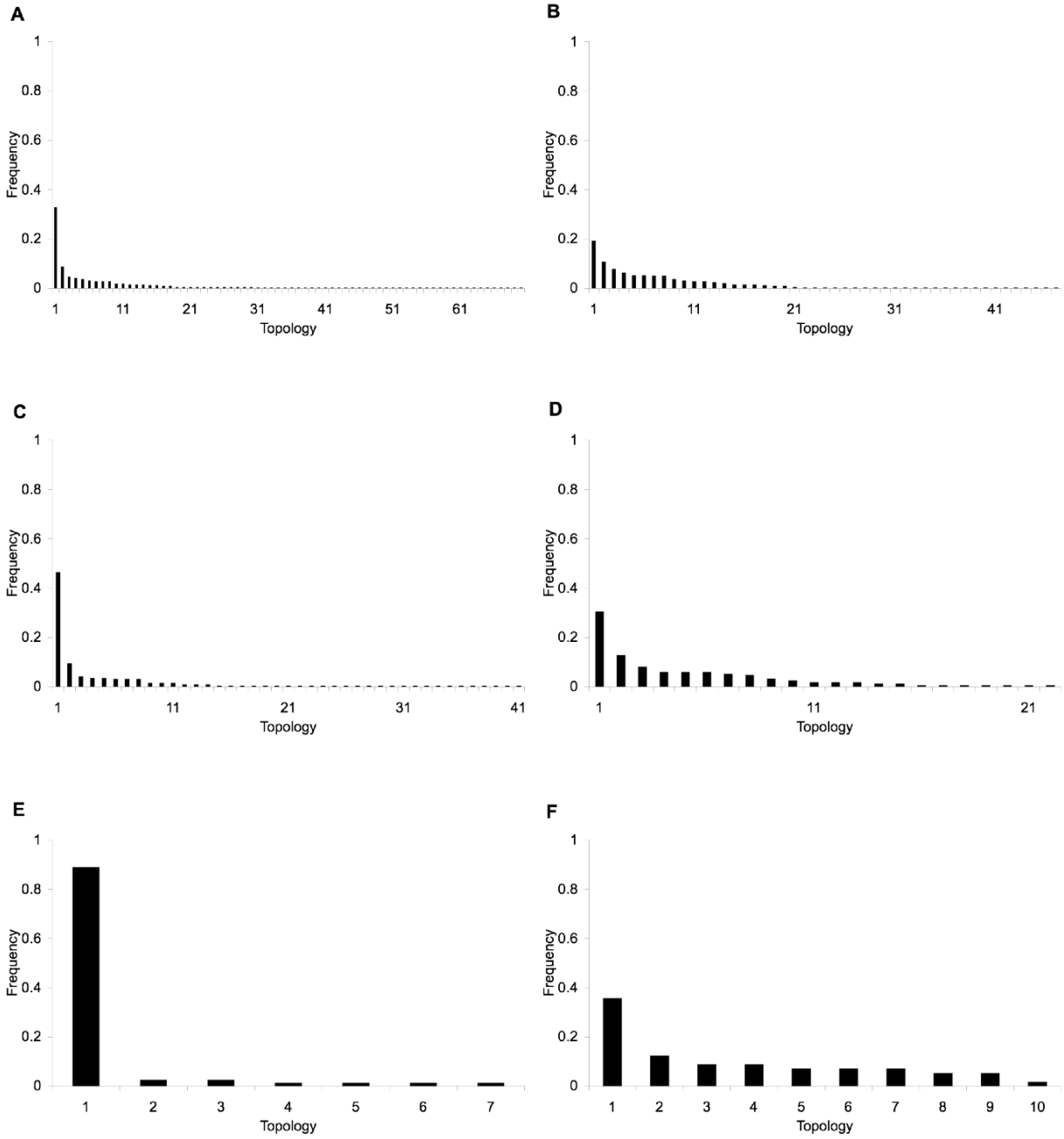


Figure 3.7. Frequency distribution of gene tree topologies. The species tree topology has the highest frequency in all panels. A. All vertebrate genes. B. All apicomplexan genes. C. Vertebrate genes passing signal filtering. D. Apicomplexan genes passing signal filtering. E. Vertebrate informative genes. F. Apicomplexan informative genes.

CHAPTER 4

CONSISTENT AND CONTRASTING PROPERTIES OF LINEAGE-SPECIFIC GENES IN THE APICOMPLEXAN PARASITES *PLASMODIUM* AND *THEILERIA*[1]

[1] Kuo, C.-H. and J.C. Kissinger. 2008. *BMC Evolutionary Biology.* 8:108.

Abstract

Background

Lineage-specific genes, the genes that are restricted to a limited subset of related organisms, may be important in adaptation. In parasitic organisms, lineage-specific gene products are possible targets for vaccine development or therapeutics when these genes are absent from the host genome.

Results

In this study, we utilized comparative approaches based on a phylogenetic framework to characterize lineage-specific genes in the parasitic protozoan phylum Apicomplexa. Genes from species in two major apicomplexan genera, *Plasmodium* and *Theileria*, were categorized into six levels of lineage specificity based on a nine-species phylogeny. In both genera, lineage-specific genes tend to have a higher level of sequence divergence among sister species. In addition, species-specific genes possess a strong codon usage bias compared to other genes in the genome. We found that a large number of genus- or species-specific genes are putative surface antigens that may be involved in host-parasite interactions. Interestingly, the two parasite lineages exhibit several notable differences. In *Plasmodium*, the (G + C) content at the third codon position increases with lineage specificity while *Theileria* shows the opposite trend. Surface antigens in *Plasmodium* are species-specific and mainly located in sub-telomeric regions. In contrast, surface antigens in *Theileria* are conserved at the genus level and distributed across the entire lengths of chromosomes.

Conclusions

Our results provide further support for the model that gene duplication followed by rapid divergence is a major mechanism for generating lineage-specific genes. The result that many lineage-specific genes are putative surface antigens supports the hypothesis that lineage-specific genes could be important in parasite adaptation. The contrasting properties between the lineage-specific genes in two major apicomplexan genera indicate that the mechanisms of generating lineage-specific genes and the subsequent evolutionary fates can differ between related parasite lineages. Future studies that focus on improving functional annotation of parasite genomes and collection of genetic variation data at within- and between-species levels will be important in facilitating our understanding of parasite adaptation and natural selection.

Background

Comparative genomics has revealed pronounced differences in gene content across species [1]. In an early analysis of eight microbial genomes, 20-56% of the genes in a genome were shown to not have high similarity to any sequence in public databases [2]. Initially these genes were referred to as orphan genes, or ORFans, because they correspond to stretches of open reading frame in bacterial genomes that have no known relationship to other sequences. As more eukaryote genome sequences become available, the term 'lineage-specific gene' is gaining in popularity because one can specify the 'lineage specificity' of a gene to describe its phylogenetic distribution [3].

Newly evolved genes may be important for adaptation and generation of diversity [4]. For example, the protozoan parasite *Cryptosporidium parvum* possesses a set of nucleotide salvage genes that are unique among all apicomplexans surveyed to date [5]. Acquisition of the nucleotide salvage pathway from a proteobacterial source as well as other sources apparently facilitated loss of genes involved in *de novo* pyrimidine biosynthesis, rendering this parasite entirely dependent on the host for both its purines and pyrimidines. Characterization of these lineage-specific genes not only leads to a better understanding of the parasite's biology but also provides a promising therapeutic target against an important parasite, since blocking the nucleotide salvage pathway can inhibit parasite growth but not harm its human host [5].

Currently, there are several hypotheses regarding the origin of lineage-specific genes. The first model invokes the process of horizontal gene transfer, in which organisms acquire genes from other distantly related species. This mechanism can create lineage-specific genes that are not shared by closely related organisms, as in the example of nucleotide salvage enzymes in *C. parvum* [5]. Previous studies have shown that horizontal gene transfer is an important force

for genome evolution in bacteria [6-8], unicellular eukaryotes [9], and multicellular eukaryotes [10].

The second model is based on gene duplication followed by rapid sequence divergence [11, 12]. Based on the observation that the sequence divergence rate is positively correlated with lineage specificity in a diverse set of organisms [3, 11-14], Alba and Castresana [12] proposed that newly duplicated genes may be released from selective constraint and accumulate mutations at a faster rate. While most of the mutations may be deleterious and lead to loss of function in one copy [15], it is also possible that one of the copies can acquire new functions and becomes a novel gene in the genome. However, whether gene duplication followed by rapid divergence is truly an important mechanism of generating lineage-specific genes is still under debate. Elhaik *et al.* [16] suggested that the correlation between divergence rate and lineage specificity may simply be an artifact, stemming from our inability to identify homologs of fast-evolving genes across distantly related taxa based on sequence similarity searches. However, a recent simulation study by Alba and Castresana [17] demonstrated that sequence similarity searches performed at the amino acid level can reliably detect fast-evolving genes due to the rate heterogeneity among sites.

In addition to the two main models discussed above, other explanations for the origin of lineage-specific genes such as *de novo* creation from non-coding sequences [18, 19], exon-shuffling [20, 21], intracellular gene transfer between organellar and nuclear genomes [9], and differential gene loss [22] also have been proposed. However, the relative importance of various forces that generate lineage-specific genes remains largely unknown.

While erroneous annotation has also been proposed as one explanation for the abundance of lineage-specific genes [23, 24], expression data [25, 26] and nucleotide substitution patterns

[24, 27] suggest that many lineage-specific genes are indeed functional and not annotation artifacts. Unfortunately, understanding the biological function of these genes is difficult due to the lack of homologs in model organisms to use for functional characterization. As a result, a large percentage of the lineage-specific genes that have been identified to date are annotated as hypothetical proteins of unknown function.

In this study, we aim to characterize the lineage-specific genes in a group of unicellular eukaryotes from the phylum Apicomplexa, including several important pathogens of humans and animals. The most infamous member of this phylum is the causative agent of malaria, *Plasmodium*, which causes more than one million human deaths per year globally [28]. Other important lineages include *Cryptosporidium* that causes cryptosporidiosis in humans and animals [29, 30], *Theileria* that causes tropical theileriosis and East Coast fever in cattle [31, 32], and *Toxoplasma* that causes toxoplasmosis in immunocompromised patients and congenitally infected fetuses [33]. The availability of genome sequences from these apicomplexan species has provided us with new and exciting opportunities to study their genome evolution. Improved knowledge of the lineage-specific genes in these important parasites can lead to a better understanding of their adaptation history and possibly identification of novel therapeutic targets.

Results

Inference of the species tree

We based our comparative analyses on a phylogenetic framework in order to infer the lineage specificity of individual genes. Among the nine species included in the data set (seven apicomplexans as well as two outgroup ciliates), we identified 83 single-copy genes that contain at least 100 alignable amino acid sites to infer the species tree (see Methods for details; a list of

these 83 genes is provided in Appendix D). Based on the concatenated alignment of these 83 genes (with 24,494 aligned amino acids sites), we infer a species tree with strong bootstrap support (Figure 4.1). This tree is consistent with our prior understanding of apicomplexan relationships based on morphology and development [34], rDNA analyses [35, 36], and multigene phylogenies [37, 38].

Phylogenetic distribution of orthologous genes

Using the species tree (Figure 4.1) as the foundation, we characterized the phylogenetic distribution of orthologous gene clusters among the apicomplexan genomes analyzed (Figure 4.2). The orthologous gene identification was performed using OrthoMCL [39] based on sequence similarity searches with an additional step of Markov Clustering [40] to improve sensitivity and specificity (see Methods for details). Our results indicated that many genes are genus-specific, ranging from approximately 30% of the genes in *Plasmodium* and *Theileria* up to about 45% in *Cryptosporidium*.

We selected *Plasmodium falciparum* and *Theileria annulata* for further investigations of lineage-specific genes. The asymmetrical topology of the species tree allows categorization of the genes in these two species into six levels of lineage specificity (Figure 4.2), yielding the highest resolution in determining the lineage specificity of a gene. The least specific genes at level 1, denoted as Pf1 for those in the *P. falciparum* genome and Ta1 for those in the *T. annulata* genome, are shared by all nine species analyzed, including two free-living ciliates; the most specific genes at level 6, denoted as Pf6 for those in the *P. falciparum* genome and Ta6 for those in the *T. annulata* genome, are species-specific. Together these six sets of genes account for 77% of annotated *P. falciparum* proteins (4,141/5,411) and 84% of annotated *T. annulata*

proteins (3,191/3,795). Genes that are shared by a non-monophyletic group (e.g., shared by *P. falciparum* and *T. annulata* but are not found in any other species) are omitted from the following analyses. Additionally, the two species pairs, *P. falciparum-P.vivax* and *T. annulata-T. parva*, may have comparable divergence times in the range of approximately 80-100 million years [41, 42] such that we can directly compare the properties of their species-specific genes. Finally, within the two focal genera, *P. falciparum* and *T. annulata* have a higher level of completeness of genome assembly than their sister species and thus are better choices for determining the chromosomal location of the lineage-specific genes.

Sequence divergence

The two *Plasmodium* species, *P. falciparum* and *P. vivax*, differ greatly in their base composition. In the coding region, *P. falciparum* has a (G + C) content of 24% while *P. vivax* has a (G + C) content of 46%. Estimates of d_N (the number of nonsynonymous substitutions per nonsynonymous site) and d_S (the number of synonymous substitutions per synonymous site) are not reliable due to the extreme AT-bias in the *P. falciparum* genome. The average d_S calculated from 4,159 *P. falciparum-P. vivax* sequence pairs is 45.7. For this reason, we quantified sequence divergence at the amino acid level based on the protein distance calculated by TREE-PUZZLE [43]. We found that the level of sequence divergence between sister taxa is positively correlated with the lineage specificity of a gene (Figure 4.3). The same trend is observed in both species-pairs. Compared to the two *Plasmodium* species, the *Theileria* species-pair has a lower level of sequence divergence. Level 6 genes are not included in the sequence divergence result because they are species-specific and have no orthologous sequence in the sister species for comparison.

We identified 1,701 genes that are single copy in both *Theileria* species and are reasonably conserved for substitution rate analysis at the nucleotide level (i.e., d_S <= 1). Consistent with the sequence divergence measured at the amino acid level, nucleotide substitution rates are higher in genes with higher lineage specificity (Table 4.2). We do not find strong evidence of any gene under positive selection (i.e., d_N/d_S ratio > 1, data not shown).

(G + C) content and relative codon bias

The average (G + C) content at the third codon position (i.e., (G+C3)) increases with lineage specificity in *P. falciparum* (Figure 4.4), suggesting that phylogenetically conserved genes are biased toward AT-rich codons in this extremely AT-rich genome. In *T. annulata*, the opposite trend is observed; genes with high lineage specificity have a lower (G + C) content at the third codon position (Figure 4.4).

We used the relative codon bias developed by Karlin *et al.* [44] to compare the differences in codon usage between different gene sets within each species (Table 4.3). In both *P. falciparum* and *T. annulata*, the level 6 (i.e., species-specific) genes exhibit a high level of deviation with regard of their codon preference compared to the other gene sets (see Methods for details). In *P. falciparum*, the average pairwise difference in all comparisons is 0.049 and the mean pairwise difference involving Pf6 genes is 0.102 (Table 4.3A). In *T. annulata*, the average pairwise difference in all comparison is 0.098 and the mean pairwise difference involving Ta6 genes is 0.183 (Table 4.3B).

Functional analyses based on annotation

As expected, most of the phylogenetically conserved genes have functional annotation or have at least one identifiable protein domain (Table 4.4). As the phylogenetic distribution of a gene becomes more restricted, it is more likely to be annotated as a hypothetical protein. Functional analysis based on available gene annotation indicates that most conserved genes (levels 1 and 2) are responsible for basic cellular processes (e.g., DNA replication, transcription, translation, etc), while most genus- and species-specific genes (levels 5 and 6) are hypothetical proteins of unknown function (data not shown). Despite the poor annotation of genus- and species-specific genes, 87% of level 5 genes and 72% of level 6 genes in *P. falciparum* have expression data available based on oligonucleotide microarrays [26]. This result suggests that most of the hypothetical proteins are real genes and not annotation artifacts.

The two focal lineages in our analysis, *Plasmodium* and *Theileria*, exhibit one interesting difference in terms of the phylogenetic distribution of surface antigens. We found that surface antigens are species-specific in *Plasmodium* and genus-specific in *Theileria*. All members of the three large surface antigen protein families in *P. falciparum* genome, including 161 rifin, 74 PfEMP1, and 35 stevor, are found in the Pf6 list and have no ortholog in *P. vivax*. Of the 163 *T. annulata* proteins that contain FAINT, a protein domain that associates with proteins exported to the host cell [31], 116 are in the Ta5 list (i.e., shared by *T. annulata* and *T. parva*) and only 28 are in the Ta6 list (i.e., specific to *T. annulata*).

In *P. falciparum* 41% of the genus-specific proteins and 62% of the species-specific proteins contain a putative signal peptide or at least one predicted transmembrane domain (Table 4.4), which suggests that these proteins may be exported to the host cell or present on the surface of the parasite or its vacuole. This result is consistent with the hypothesis that lineage-specific

genes in apicomplexan parasites are likely to be involved in host-parasite interactions and thus, potentially adaptation.

Chromosomal location

Analysis of chromosomal location demonstrated that most species-specific genes in *P. falciparum* are located near chromosome ends (see Figure 4.5 for one example chromosome and Appendix E for all 14 chromosomes). In *T. annulata* (see Figure 4.6 for one example chromosome and Appendix F for all four chromosomes), we observed a similar pattern that the regions adjacent to chromosome ends are devoid of the phylogenetically conserved genes (cf. Figures 4.5B and 4.6B). However, unlike the pattern found in *P. falciparum*, most of the species-specific genes in *T. annulata* (i.e., Ta6) are distributed across the entire length of chromosomes and are not enriched in the regions adjacent to chromosome ends (cf. Figures 4.5A and 4.6A).

To quantify the pattern of gene distribution on chromosomes, we calculated the distance of each gene to the nearest chromosome end. For each set of genes (levels 1 through 6 in each species), we utilized (1) the average distance to the nearest chromosome end and (2) the minimal distance to the nearest chromosome end (i.e., the minimal found in a given gene set) for this analysis. In *P. falciparum*, the average distance scales with chromosome size and the species-specific genes (i.e., Pf6) are closer to chromosome ends (Figure 4.7A). In contrast, minimal distance does not scale with chromosome size (Figure 4.7B). For all chromosomes, the minimal distances of phylogenetically conserved genes (i.e., Pf1 through Pf4) from the chromosome ends are larger than 50-100kb. This result indicates that the regions that are occupied exclusively by genus- and species-specific genes are proportionally larger in smaller chromosomes. Consistent with this observation, three of the smallest chromosomes in *P. falciparum* (i.e., MAL1, MAL2,

and MAL4) have many more species-specific genes than random expectation (Chi-square test d.f. = (6 gene sets -1) * (14 chromosomes - 1) = 65, P-value = 1e-12).

In *T. annulata*, genes with different levels of lineage specificity have similar average distances to chromosome ends (Figure 4.7C). This result corroborates the visual pattern in Figure 4.6A that species-specific genes are distributed across the entire length of a chromosome, in contrast to the clustering near chromosome ends observed in *P. falciparum* (Figure 4.5A). For all four chromosomes in *T. annulata*, the regions that are adjacent to chromosome ends and devoid of phylogenetically conserved genes (i.e., Ta1 through Ta4) are approximately 20-40kb (Figure 4.7D), a distance smaller than in *P. falciparum*. Unlike the pattern found in *P. falciparum* in which species-specific genes are closer to chromosome ends than genus-specific genes, genus- and species-specific genes in *T. annulata* (i.e., Ta5 and Ta6) have similar minimal distances in all four chromosomes (Figure 4.7D).

In both *P. falciparum* and *T. annulata*, genes located near chromosome ends have a higher level of sequence divergence relative to its ortholog in the sister species at the amino acid level (Figure 4.8). This trend is observed in genes with different levels of lineage specificity and is stronger in *T. annulata*.

Discussion

We identified a pattern in which lineage-specific genes have a higher level of sequence divergence among sister species in a group of important protozoan parasites. This result is consistent with previous studies in bacteria [13], fungi [3], and animals [11, 12, 14]. Now we further confirm that this pattern also holds true in a protistan phylum, suggesting that it may be universal across much of the tree-of-life. Results from functional analyses agree with our

intuitive expectation that conserved genes are involved in basic cellular functionalities and are well annotated. A large number of the lineage-specific genes (at the species level in *Plasmodium* and the genus level in *Theileria*) are found to be putative surface antigens that the parasites use to interact with their hosts. This result supports the hypothesis that lineage-specific genes may be important in adaptation [4]. In addition, the physical distance of a gene to the nearest chromosome end is correlated with the level of sequence divergence.

We found three contrasting properties of lineage-specific genes between two major apicomplexan lineages. First, families of surface antigens are species-specific in *Plasmodium* but genus-specific in *Theileria*. Second, most of the species-specific genes are located in sub-telomeric regions in *P. falciparum* but no such pattern exists in *T. annulata.* Third, the (G + C) content at the third codon position increases with lineage specificity in *P. falciparum* but decreases in *T. annulata*. Taken together, these results suggest that the mechanisms of generating lineage-specific genes and their subsequent evolutionary fates differ between apicomplexan parasite lineages.

Gene content evolution

All apicomplexan species analyzed have small genomes compared to the free-living out-group. This result is consistent with comparative genomic analyses conducted in other pathogenic bacteria and eukaryotes; extreme genome reduction is a common theme in the genome evolution of these organisms [45].

A large proportion of the genes in apicomplexans are genus-specific (Figure 4.2). One parsimonious explanation for this observation is that each lineage acquired a new set of genes during its evolutionary history. An alternative explanation invokes differential loss among

lineages when evolving from a free-living ancestor with a relatively large genome. We found that 23% of the protein coding genes in *P. falciparum* and 16% in *T. annulata* have a complex phylogenetic distribution pattern and do not fit into a simple single gain/loss model. These results suggest that some ancestral genes in the apicomplexans may have experienced multiple independent losses during their evolutionary history. Further investigation is necessary to distinguish true gene gains from differential retention of ancestral genes.

Comparison of genes with different levels of lineage specificity

Consistent with previous studies in bacteria [13], fungi [3], and animals [11, 12, 14], we observed a pattern in which sequence divergence is higher in genes with a higher level of lineage specificity. One explanation is that phylogenetically conserved genes are often involved in fundamental cellular processes (see Results). These genes are likely to be under purifying selection that constrains the rate of sequence divergence. In support of this hypothesis, we observe that the mean d_N/d_S ratio among the level 1 genes in *Theileria* is only 0.07 (Table 4.2), indicating an extremely low rate of nonsynonymous substitution relative to synonymous substitution.

Based on the hypothesis that lineage-specific genes are often involved in adaptation [4], such as invasion of hosts or evasion of the immune responses, lineage-specific genes may be under positive selection and have a faster rate of sequence divergence. Our data is suggestive in this regard, as genus-specific genes exhibit higher sequence divergence than genes with lower levels of lineage specificity. Unfortunately we cannot directly test the hypothesis that lineage-specific genes are more likely to be under positive selection using the d_N/d_S ratio data. The level of sequence divergence is too high in both species pairs for such analysis. Practically all of the

genes from the *Plasmodium* pair and approximately 1,000 genes from the *Theileria* pair (i.e., more than a quarter of the gene repertoire) have a d_S estimate that is larger than one. Under this high level of sequence divergence, we cannot confidently estimate the substitution rate due to saturation. Better detection of positive selection in these genes requires data on genetic variation at within- and between-species levels [46, 47].

Codon bias analyses indicate that species-specific genes have a different codon preference compared to other genes in the same genome, whereas the genes with lower levels of lineage specificity are relatively similar to each other (Table 4.3). It is possible that species-specific genes are relatively young and have yet to adapt to the codon usage pattern of the genome. Support for this hypothesis provided by the observation that the (G + C) content at the third codon position is much lower in the phylogenetically conserved genes in *P. falciparum* (Figure 4.4), suggesting that these 'older' genes are more biased toward GC-poor codons in this AT-rich genome. Alternatively, some species-specific genes may be subject to a different pattern of selection and thus possess different codon preference.

For the lineage-specific genes at the genus and species level that have functional annotations, many are known surface antigens. Because surface antigens are used by the parasites to interact with their hosts [48], such as adhesion to the cell surface or evasion of the host immune response, this result supports the hypothesis that (at least some) lineage-specific genes are involved in host-parasite interactions and have facilitated lineage-specific adaptation. Interestingly, surface antigens are species-specific in *Plasmodium*, but are genus-specific in *Theileria*. In addition, 62% of *P. falciparum*-specific genes contain a putative signal peptide or at least one predicted transmembrane domain. This result is consistent with one previous study that compared *P. falciparum* with three other *Plasmodium* species that cause rodent malaria [49]. Of

the 168 *P. falciparum*-specific genes identified in this previous study that are not located in sub-telomeric regions, 68% are predicted to be exported to the surface of the parasites or the infected host cells.

Comparison between Plasmodium and Theileria

Previous studies suggest that the two focal species pairs have similar divergence times. The two *Plasmodium* species diverged about 80-100 million years ago [41] and the two *Theileria* species diverged about 82 million years ago [42]. Our results indicate that sequence divergence is much higher between the two *Plasmodium* species (Figures 1 and 3). This may be caused by the difference in nucleotide composition, since *P. falciparum* has a GC content of 24% while *P. vivax* has a GC content of 46% in the coding region. Bias in nucleotide composition has been shown to change codon usage and amino acid composition [50]. Alternatively, it is also possible that the divergence time between *T. annulata* and *T. parva* was overestimated because it was based on a simplified assumption that the synonymous substitution rate in *Theileria* is similar to that in *Plasmodium* [42].

In both *P. falciparum* and *T. annulata*, the sub-telomeric regions contain exclusively genus- or species-specific genes. Interestingly, the physical size of these regions is not correlated with chromosome size. This observation indicates that these regions are proportionally larger in smaller chromosomes and helps explains the pattern that the three small chromosomes in *P. falciparum* have many more species-specific genes than predicted by random expectations (see Results). In addition, genes that are located near a chromosome end have a higher level of sequence divergence in both species, regardless of their lineage specificity (Figure 8). The high evolutionary rates in sub-telomeric regions are shared by many eukaryotic lineages; high rates of

inter-chromosomal recombination, local duplication, and segmental rearrangement have been reported in organisms including humans [51], yeasts [52], and plants [53].

Given the high rates of evolution in sub-telomeric regions, it may be advantageous for pathogens to have their surface antigen genes located in these evolutionary hotspots to facilitate the generation of antigenic diversity. Consistent with this hypothesis, many micro-parasites have large gene families that encode surface antigens in sub-telomeric regions (reviewed in [54]). The best-studied example is the causative agent of African trypanosomiasis, *Trypanosoma brucei*. The *vsg* gene family in *T. brucei* encodes variant surface glycoproteins (VSG) that form a dense coat on the outside of the parasite. In the bloodstream stage, *T. brucei* sequentially expresses different members of the *vsg* gene family, one at a time, to generate antigenic variation [55]. The positioning of *vsg* genes in the genome is tightly linked to regulation of expression; the actively expressed *vsg* is duplicated into one of the bloodstream expression sites located in the sub-telomeric regions (reviewed in [56, 57]). This homologous recombination process which involves loci that are not positional alleles is hypothesized to be important in generating genetic diversity within the gene family [54]. Although the genes encoding surface antigens in *P. falciparum* are not known to be duplicated into specific expression sites as observed in *T. brucei*, the clustering of these genes in sub-telomeric regions can facilitate inter-chromosomal recombination that increases antigenic variation [58].

We found that most of the surface antigen genes in *P. falciparum* are located in sub-telomeric regions, as previously noted [28]. Several studies have established the importance of genome location in the generation and maintenance of antigenic variation in *P. falciparum* [58, 59]. The surface antigen PfEMP1 possessed by *P. falciparum* is exported to the cell surface of infected erythrocytes. PfEMP1 can remove infected erythrocytes from blood circulation by

cellular adherence to microvascular endothelial cells and avoid spleen-dependent killing [60]. The study on genetic structuring suggested that the approximately 60 copies of *var* genes (which encode PfEMP1) in the *P. falciparum* genome can be divided into three functionally diverged groups with two in sub-telomeric regions and one close to the centers of chromosomes [59]. Furthermore, the recombination rate is found to be high among members in the same functional group but low for members belonging to different groups. This recombinational hierarchy may facilitate the generation of genetic diversity within a group and promote specialization between different groups. Experimental evidence suggests that the clustering of *var* genes in the sub-telomeric regions is important in the epigenetic regulation of gene expression in *P. falciparum* [61, 62].

Given the generality of association between surface antigen genes and sub-telomeric regions in micro-parasites, it is interesting to see that *T. annulata* appears to be an exception to this rule. This finding may provide an explanation for the difference in host range between the two apicomplexan lineages. Because a large percentage of surface antigen genes in *Plasmodium* are located in sub-telomeric regions, the generation of antigenic variation may be faster in *Plasmodium* than in *Theileria*. Our results indicate that gene families encoding surface antigens in *Plasmodium* are highly diverged between species within the genus, whereas the two *Theileria* species still share most of their surface antigens and the genes encoding them are distributed across the entire lengths of chromosomes. For this reason, *Plasmodium* may be able to adapt to new host species at a faster rate, resulting in its much wider host range compared to *Theileria*; *Plasmodium spp.* can infect mammals, birds, and reptiles, whereas *Theileria spp.* are limited to ruminants [34].

Conclusions

Our results agree with previous observations in other organisms that lineage-specific genes have a higher level of sequence divergence compared to phylogenetically conserved genes. In addition, two major apicomplexan lineages may have different mechanisms for generating or retaining species-specific genes. Because many lineage-specific genes in these parasites are surface antigens that interact with the host, future investigations on genome evolution in these parasites may facilitate the identification of new therapeutic or vaccine targets. Future studies that focus on improving functional annotation of parasite genomes and the collection of genetic variation data at different phylogenetic levels will be important in our understanding of parasite adaptation and natural selection.

Methods

Data source and orthologous gene identification

The data sources of the annotated proteins are listed in Table 4.1. Protein domain identification was performed with HMMPFAM [63] (version 20.0). Transmembrane domain prediction [28] and gene expression data [26] of annotated *Plasmodium falciparum* genes were downloaded from PlasmoDB [64] (Release 5.3).

Orthologous gene clusters were identified using OrthoMCL [39] (version 1.3, April 10, 2006) with default parameter settings. 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 [40] to improve sensitivity and specificity. We used WU-BLAST [65] (version 2.0) for the all-against-all BLASTP similarity search step with the e-value cutoff set to 1e-15.

Phylogenetic inference

Based on the orthologous gene clustering result, we identified genes that are shared by all nine species to infer the species tree. Orthologous gene clusters that contain more than one gene from any given species were removed to avoid the complications introduced by paralogous gene in phylogenetic inference. Of the 768 orthologous gene clusters that are shared by all nine species (Figure 4.2), 154 clusters were single-copy in all species. For each gene, CLUSTALW [66] (version 1.83) was used for multiple sequence alignment. We enabled the 'tossgaps' option to ignore gaps when constructing the guide tree and used the default settings for all other parameters. The alignments produced by CLUSTALW were filtered by GBLOCKS [67] (version 0.91b) to remove regions that contain gaps or are highly divergent. Individual genes that had less than 100 aligned amino acid sites (33/154) or contained identical sequences from different taxa (38/154) after GBLOCKS filtering were eliminated from further analysis. We concatenated the alignments from the remaining 83 genes (with a total of 24,494 aligned amino acid sites) and utilized PHYML [68] to infer the species tree based on the maximum likelihood method. We used PHYML to estimate the proportion of invariable sites and the gamma distribution parameter (with eight substitution categories). The substitution model was set to JTT [69] and we enabled the optimization options for tree topology, branch lengths, and rate parameters. To estimate the level of support on each internal branch, we performed 100 non-parametric bootstrap samplings.

Quantification of sequence divergence

The nonsynonymous and synonymous substitution rates at the nucleotide level (i.e., d_N and d_S) were estimated using CODEML in the PAML package [70]. We performed pairwise sequence alignment at the amino acid level using CLUSTALW [66] with default parameters for

all orthologous genes that are single copy in both *Plasmodium* species or both *Theileria* species. The protein alignments were converted into the corresponding nucleotide alignments using NAL2PAL [71] (version 12). All gap positions were removed from the alignments before the substitution rate estimation by CODEML. To avoid problems of inaccurate rate estimation caused by saturation, we excluded sequences with a synonymous substitution rate (d_S) that is greater than one.

To quantify the level of sequence divergence at the amino acid level, we used TREE-PUZZLE [43] to calculate the protein distance between orthologs in sister species. The parameters were set to the JTT substitution model [69], mixed model of rate heterogeneity with one invariable and eight Gamma rate categories, and the exact and slow parameter estimation. Orthologous sequences were first aligned using CLUSTALW [66] followed by a filtering step using GBLOCKS [67] to remove gaps and highly divergent regions before the calculation of protein distance. Five sequences (PFA0650w, PFD0105c, PFL0060w, and PFD1140w from *P. falciparum* and TA18345 from *T. annulata*) that were not reliably aligned to their ortholog in the sister species were excluded from this analysis.

Calculation of relative codon bias

The relative codon bias between sets of genes in the two focal species, *P. falciparum* and *T. annulata,* was calculated based on the method developed by Karlin et al. [44]. Briefly, the method considers two sets of genes, one focal set and one reference set, and calculates the difference in relative frequency of codon family that encode the same amino acid between the two sets. The theoretical maximum of the difference between two sets of genes is 2.000, but the empirical values based on biological data generally range from 0.050 to 0.300 [44, 72, 73]. This

measurement is different from the conventional codon adaptation index (CAI) developed by Sharp and Li [74], in which a set of highly expressed genes is always used as the reference set. We choose the relative codon bias to measure codon preference because it can provide a better resolution under certain conditions. For example, two sets of weakly expressed genes may have similar values of codon adaptation index but still possess vastly different codon preferences.

Visualization and quantification of chromosomal location

GBROWSE [75] was used for visualization of gene distribution on chromosomes. To quantify the pattern of chromosomal location, we calculated the distance of each gene to the nearest chromosome end. For example, the *P. falciparum* gene PF10_0023 on chromosome MAL10 (physical size is 1,694,445 bp) starts at position 99,380 and ends at 100,362. Its distance to the nearest chromosome end was calculated as 99,380 - 1 = 99,379 bp. For gene PF10_0369 on the same chromosome that starts at 1,493,991 and ends at 1,496,955, its distance to the nearest chromosome end was calculated as 1,694,445 - 1,496,955 = 197,490 bp. The orientation of a gene (i.e., whether it is on the '+' strand or the '-' strand) is ignored for distance calculation.

Authors' contributions

CHK developed the concept of this study, performed the analysis, and wrote the manuscript. JCK provided supervision, feedback, and comments on the manuscript. All authors have read and approved the final manuscript.

Acknowledgements

CHK was supported by a 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 JCK. The Institute of Bioinformatics and the Research Computing Center at the University of Georgia provided computation resources. P. Brunk, F. Chen, J. Felsenstein, M. Heiges, J. Mrazek, A. Oliveira, E. Robinson, and H. Wang provided valuable assistance on the use of computer hardware and software. D. Promislow, J. Bennetzen, D. Hall, J. Linder, J. Moorad, B. Striepen, and four anonymous reviewers provided helpful comments that improved the manuscript. We thank the J. Craig Venter Institute for providing pre-publication access to the genome sequence data of *Plasmodium vivax* and *Toxoplasma gondii*. The US Department of Defense, the National Institute of Allergy and Infectious Disease, and the Burroughs Wellcome Fund provided funding for the genome sequencing project of *Plasmodium vivax* and *Toxoplasma gondii*.

References

1. Rubin GM, Yandell MD, Wortman JR, Gabor Miklos GL, Nelson CR, Hariharan IK, Fortini ME, Li PW, Apweiler R, Fleischmann W *et al*: **Comparative genomics of the eukaryotes**. *Science* 2000, **287**(5461):2204-2215.

2. Fischer D, Eisenberg D: **Finding families for genomic ORFans**. *Bioinformatics* 1999, **15**(9):759 - 762.

3. Cai J, Woo P, Lau S, Smith D, Yuen K-y: **Accelerated evolutionary rate may be responsible for the emergence of lineage-specific genes in Ascomycota**. *J Mol Evol* 2006, **63**(1):1-11.

4. Wilson GA, Bertrand N, Patel Y, Hughes JB, Feil EJ, Field D: **Orphans as taxonomically restricted and ecologically important genes**. *Microbiology* 2005, **151**(8):2499-2501.

5. Striepen B, Pruijssers AJP, Huang JL, Li C, Gubbels MJ, Umejiego NN, Hedstrom L, Kissinger JC: **Gene transfer in the evolution of parasite nucleotide biosynthesis**. *Proc Natl Acad Sci USA* 2004, **101**(9):3154-3159.

6. Gogarten JP, Townsend JP: **Horizontal gene transfer, genome innovation and evolution**. *Nat Rev Microbiol* 2005, **3**(9):679-687.

7. Ochman H, Lawrence JG, Groisman EA: **Lateral gene transfer and the nature of bacterial innovation**. *Nature* 2000, **405**(6784):299-304.

8. Lerat E, Daubin V, Ochman H, Moran NA: **Evolutionary origins of genomic repertoires in bacteria**. *PLoS Biol* 2005, **3**:e130.

9. Huang JL, Mullapudi N, Sicheritz-Ponten T, Kissinger JC: **A first glimpse into the pattern and scale of gene transfer in the Apicomplexa**. *Int J Parasitol* 2004, **34**(3):265-274.

10. Hotopp JCD, Clark ME, Oliveira DCSG, Foster JM, Fischer P, Torres MCM, Giebel JD, Kumar N, Ishmael N, Wang S *et al*: **Widespread lateral gene transfer from intracellular bacteria to multicellular eukaryotes**. *Science* 2007, **317**(5845):1753-1756.

11. Domazet-Loso T, Tautz D: **An evolutionary analysis of orphan genes in *Drosophila***. *Genome Res* 2003, **13**(10):2213 - 2219.

12. Alba MM, Castresana J: **Inverse relationship between evolutionary rate and age of mammalian genes**. *Mol Biol Evol* 2005, **22**(3):598-606.

13. Daubin V, Ochman H: **Bacterial genomes as new gene homes: the genealogy of ORFans in *E. coli***. *Genome Res* 2004, **14**(6):1036 - 1042.

14. Wang W, Zheng H, Yang S, Yu H, Li J, Jiang H, Su J, Yang L, Zhang J, McDermott J *et al*: **Origin and evolution of new exons in rodents**. *Genome Res* 2005, **15**(9):1258 - 1264.

15. Kellis M, Birren BW, Lander ES: **Proof and evolutionary analysis of ancient genome duplication in the yeast *Saccharomyces cerevisiae***. *Nature* 2004, **428**(6983):617-624.

16. Elhaik E, Sabath N, Graur D: **The "Inverse relationship between evolutionary rate and age of mammalian genes" is an artifact of increased genetic distance with rate of evolution and time of divergence**. *Mol Biol Evol* 2006, **23**:1 - 3.

17. Alba MM, Castresana J: **On homology searches by protein Blast and the characterization of the age of genes**. *BMC Evol Biol* 2007, **7**(1):53.

18. Levine MT, Jones CD, Kern AD, Lindfors HA, Begun DJ: **Novel genes derived from noncoding DNA in *Drosophila melanogaster* are frequently X-linked and exhibit testis-biased expression**. *Proc Natl Acad Sci USA* 2006, **103**(26):9935-9939.

19. Chen S-T, Cheng H-C, Barbash DA, Yang H-P: **Evolution of *hydra,* a recently evolved testis-expressed gene with nine alternative first exons in *Drosophila melanogaster***. *PLoS Genet* 2007, **3**(7):e107.

20. Patthy L: **Genome evolution and the evolution of exon-shuffling -- a review**. *Gene* 1999, **238**(1):103-114.

21. Moran JV, DeBerardinis RJ, Kazazian HH, Jr.: **Exon shuffling by L1 retrotransposition**. *Science* 1999, **283**(5407):1530-1534.

22. Blomme T, Vandepoele K, De Bodt S, Simillion C, Maere S, Van de Peer Y: **The gain and loss of genes during 600 million years of vertebrate evolution**. *Genome Biol* 2006, **7**(5):R43.

23. Schmid KJ, Aquadro CF: **The evolutionary analysis of "orphans" from the Drosophila genome identifies rapidly diverging and incorrectly annotated genes**. *Genetics* 2001, **159**(2):589-598.

24. Ochman H: **Distinguishing the ORFs from the ELFs: short bacterial genes and the annotation of genomes**. *Trends Genet* 2002, **18**(7):335-337.

25. Rosenow C, Saxena RM, Durst M, Gingeras TR: **Prokaryotic RNA preparation methods useful for high density array analysis: comparison of two approaches**. *Nucleic Acids Res* 2001, **29**(22):e112-.

26. Le Roch KG, Zhou Y, Blair PL, Grainger M, Moch JK, Haynes JD, De la Vega P, Holder AA, Batalov S, Carucci DJ *et al*: **Discovery of gene function by expression profiling of the malaria parasite life cycle**. *Science* 2003, **301**(5639):1503-1508.

27. Nekrutenko A, Makova KD, Li W-H: **The KA/KS ratio test for assessing the protein-coding potential of genomic regions: an empirical and simulation study**. *Genome Res* 2002, **12**(1):198-202.

28. Gardner MJ, Hall N, Fung E, White O, Berriman M, Hyman RW, Carlton JM, Pain A, Nelson KE, Bowman S *et al*: **Genome sequence of the human malaria parasite *Plasmodium falciparum***. *Nature* 2002, **419**(6906):498-511.

29. Abrahamsen MS, Templeton TJ, Enomoto S, Abrahante JE, Zhu G, Lancto CA, Deng M, Liu C, Widmer G, Tzipori S *et al*: **Complete genome sequence of the Apicomplexan, *Cryptosporidium parvum***. *Science* 2004, **304**:441-445.

30. Xu P, Widmer G, Wang Y, Ozaki LS, Alves JM, Serrano MG, Puiu D, Manque P, Akiyoshi D, Mackey AJ *et al*: **The genome of *Cryptosporidium hominis***. *Nature* 2004, **431**(7012):1107-1112.

31. Pain A, Renauld H, Berriman M, Murphy L, Yeats CA, Weir W, Kerhornou A, Aslett M, Bishop R, Bouchier C *et al*: **Genome of the host-cell transforming parasite *Theileria annulata* compared with *T. parva***. *Science* 2005, **309**(5731):131-133.

32. Gardner MJ, Bishop R, Shah T, de Villiers EP, Carlton JM, Hall N, Ren Q, Paulsen IT, Pain A, Berriman M *et al*: **Genome sequence of *Theileria parva*, a bovine pathogen that transforms lymphocytes**. *Science* 2005, **309**(5731):134-137.

33. Montoya JG, Liesenfeld O: **Toxoplasmosis**. *Lancet* 2004, **363**(9425):1965-1976.

34. Lee J, Leedale G, Bradbury P: **An Illustrated Guide to the Protozoa**, vol. 1, 2nd edn. Lawrence, KS, USA: Society of Protozoologists; 2000.

35. Escalante A, Ayala F: **Evolutionary origin of *Plasmodium* and other Apicomplexa based on rRNA genes**. *Proc Natl Acad Sci USA* 1995, **92**(13):5793-5797.

36. Morrison DA, Ellis JT: **Effects of nucleotide sequence alignment on phylogeny estimation: A case study of 18S rDNAs of Apicomplexa**. *Mol Biol Evol* 1997, **14**(4):428-441.

37. Douzery EJP, Snell EA, Bapteste E, Delsuc F, Philippe H: **The timing of eukaryotic evolution: Does a relaxed molecular clock reconcile proteins and fossils?** *Proc Natl Acad Sci USA* 2004, **101**(43):15386-15391.

38. Philippe H, Snell EA, Bapteste E, Lopez P, Holland PWH, Casane D: **Phylogenomics of eukaryotes: Impact of missing data on large alignments**. *Mol Biol Evol* 2004, **21**(9):1740-1752.

39. Li L, Stoeckert CJ, Roos DS: **OrthoMCL: Identification of ortholog groups for eukaryotic genomes**. *Genome Res* 2003, **13**(9):2178-2189.

40. Van Dongen S: **Graph clustering by flow simulation**. *PhD thesis.* University of Utrecht; 2000.

41. Perkins SL, Schall JJ: **A molecular phylogeny of malarial parasites recovered from cytochrome b gene sequences**. *J Parasitol* 2002, **88**(5):972-978.

42. Roy SW, Penny D: **Large-scale intron conservation and order-of-magnitude variation in intron loss/gain rates in apicomplexan evolution**. *Genome Res* 2006, **16**(10):1270-1275.

43. Schmidt HA, Strimmer K, Vingron M, von Haeseler A: **TREE-PUZZLE: maximum likelihood phylogenetic analysis using quartets and parallel computing**. *Bioinformatics* 2002, **18**(3):502-504.

44. Karlin S, Mrazek J, Campbell AM: **Codon usages in different gene classes of the *Escherichia coli* genome**. *Mol Microbiol* 1998, **29**(6):1341-1355.

45. Lawrence JG: **Common themes in the genome strategies of pathogens**. *Curr Opin Genet Dev* 2005, **15**(6):584-588.

46. Kidgell C, Volkman SK, Daily J, Borevitz JO, Plouffe D, Zhou Y, Johnson JR, Le Roch KG, Sarr O, Ndir O *et al*: **A systematic map of genetic variation in *Plasmodium falciparum***. *PLoS Pathog* 2006, **2**(6):e57.

47. Barry AE, Leliwa-Sytek A, Tavul L, Imrie H, Migot-Nabias F, Brown SM, McVean GAV, Day KP: **Population genomics of the immune evasion (*var*) genes of *Plasmodium falciparum***. *PLoS Pathog* 2007, **3**(3):e34.

48. Bull PC, Berriman M, Kyes S, Quail MA, Hall N, Kortok MM, Marsh K, Newbold CI: ***Plasmodium falciparum* variant surface antigen expression patterns during malaria**. *PLoS Pathog* 2005, **1**(3):e26.

49. Kooij TWA, Carlton JM, Bidwell SL, Hall N, Ramesar J, Janse CJ, Waters AP: **A *Plasmodium* whole-genome synteny map: indels and synteny breakpoints as foci for species-specific genes**. *PLoS Pathog* 2005, **1**(4):e44.

50. Foster P, Jermiin L, Hickey D: **Nucleotide composition bias affects amino acid content in proteins coded by animal mitochondria**. *J Mol Evol* 1997, **44**(3):282-288.

51. Linardopoulou EV, Williams EM, Fan Y, Friedman C, Young JM, Trask BJ: **Human subtelomeres are hot spots of interchromosomal recombination and segmental duplication**. *Nature* 2005, **437**(7055):94-100.

52. Ricchetti M, Dujon B, Fairhead C: **Distance from the chromosome end determines the efficiency of double strand break repair in subtelomeres of haploid yeast**. *Journal of Molecular Biology* 2003, **328**(4):847-862.

53. Kuo H-F, Olsen KM, Richards EJ: **Natural variation in a subtelomeric region of *Arabidopsis*: implications for the genomic dynamics of a chromosome end**. *Genetics* 2006, **173**(1):401-417.

54. Barry JD, Ginger ML, Burton P, McCulloch R: **Why are parasite contingency genes often associated with telomeres?** *Int J Parasitol* 2003, **33**(1):29-45.

55. Cross GAM, Wirtz LE, Navarro M: **Regulation of vsg expression site transcription and switching in *Trypanosoma brucei***. *Molecular and Biochemical Parasitology* 1998, **91**(1):77-91.

56. Navarro M, Penate X, Landeira D: **Nuclear architecture underlying gene expression in *Trypanosoma brucei***. *Trends Microbiol* 2007, **15**(6):263-270.

57. Dreesen O, Li B, Cross GAM: **Telomere structure and function in trypanosomes: a proposal**. *Nat Rev Microbiol* 2007, **5**(1):70-75.

58. Freitas-Junior LH, Bottius E, Pirrit LA, Deitsch KW, Scheidig C, Guinet F, Nehrbass U, Wellems TE, Scherf A: **Frequent ectopic recombination of virulence factor genes in telomeric chromosome clusters of *P. falciparum***. *Nature* 2000, **407**(6807):1018-1022.

59. Kraemer SM, Smith JD: **Evidence for the importance of genetic structuring to the structural and functional specialization of the *Plasmodium falciparum var* gene family**. *Mol Microbiol* 2003, **50**(5):1527-1538.

60. Baruch DI, Gormley JA, Ma C, Howard RJ, Pasloske BL: ***Plasmodium falciparum*** **erythrocyte membrane protein 1 is a parasitized erythrocyte receptor for adherence to CD36, thrombospondin, and intercellular adhesion molecule**. *Proc Natl Acad Sci USA* 1996, **93**(8):3497-3502.

61. Duraisingh MT, Voss TS, Marty AJ, Duffy MF, Good RT, Thompson JK, Freitas-Junior LH, Scherf A, Crabb BS, Cowman AF: **Heterochromatin silencing and locus repositioning linked to regulation of virulence genes in *Plasmodium falciparum***. *Cell* 2005, **121**(1):13-24.

62. Scherf A, Figueiredo LM, Freitas-Junior LH: ***Plasmodium*** **telomeres: a pathogen's perspective**. *Curr Opin Microbiol* 2001, **4**(4):409-414.

63. Bateman A, Coin L, Durbin R, Finn RD, Hollich V, Griffiths-Jones S, Khanna A, Marshall M, Moxon S, Sonnhammer ELL *et al*: **The Pfam protein families database**. *Nucleic Acids Res* 2004, **32**:D138-D141.

64. Bahl A, Brunk B, Crabtree J, Fraunholz MJ, Gajria B, Grant GR, Ginsburg H, Gupta D, Kissinger JC, Labo P *et al*: **PlasmoDB: the *Plasmodium* genome resource. A database integrating experimental and computational data**. *Nucleic Acids Res* 2003, **31**(1):212-215.

65. **WU-BLAST** [http://blast.wustl.edu/]

66. Thompson JD, Higgins DG, Gibson TJ: **CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, positions-specific gap penalties and weight matrix choice**. *Nucleic Acids Res* 1994, **22**:4673 - 4680.

67. Castresana J: **Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis**. *Mol Biol Evol* 2000, **17**(4):540-552.

68. Guindon S, Gascuel O: **A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood**. *Syst Biol* 2003, **52**:696 - 704.

69. Jones DT, Taylor WR, Thornton JM: **The rapid generation of mutation data matrices from protein sequences**. *Comput Appl Biosci* 1992, **8**(3):275-282.

70. Yang Z: **PAML 4: phylogenetic analysis by maximum likelihood**. *Mol Biol Evol* 2007, **24**(8):1586-1591.

71. Suyama M, Torrents D, Bork P: **PAL2NAL: robust conversion of protein sequence alignments into the corresponding codon alignments**. *Nucleic Acids Res* 2006, **34**(suppl_2):W609-612.

72. Karlin S, Mrazek J: **Predicted highly expressed genes of diverse prokaryotic genomes**. *J Bacteriol* 2000, **182**(18):5238-5250.

73. Karlin S, Campbell AM, Mrazek J: **Comparative DNA analysis across diverse genomes**. *Annu Rev Genet* 1998, **32**:185-225.

74. Sharp PM, Li WH: **The codon adaptation index - a measure of directional synonymous codon usage bias, and its potential applications**. *Nucleic Acids Res* 1987, **15**(3):1281-1295.

75. Stein LD, Mungall C, Shu S, Caudy M, Mangone M, Day A, Nickerson E, Stajich JE, Harris TW, Arva A *et al*: **The generic genome browser: a building block for a model organism system database**. *Genome Res* 2002, **12**(10):1599-1610.

76. Heiges M, Wang HM, Robinson E, Aurrecoechea C, Gao X, Kaluskar N, Rhodes P, Wang S, He CZ, Su YQ *et al*: **CryptoDB: a *Cryptosporidium* bioinformatics resource update**. *Nucleic Acids Res* 2006, **34**:D419-D422.

77. Hertz-Fowler C, Peacock CS, Wood V, Aslett M, Kerhornou A, Mooney P, Tivey A, Berriman M, Hall N, Rutherford K *et al*: **GeneDB: a resource for prokaryotic and eukaryotic organisms**. *Nucleic Acids Res* 2004, **32**(suppl_1):D339-343.

78. **J. Craig Venter Institute** [http://jcvi.org/]

79. Gajria B, Bahl A, Brestelli J, Dommer J, Fischer S, Gao X, Heiges M, Iodice J, Kissinger JC, Mackey AJ *et al*: **ToxoDB: an integrated *Toxoplasma gondii* database resource**. *Nucleic Acids Res* 2007:gkm981.

80. Aury J-M, Jaillon O, Duret L, Noel B, Jubin C, Porcel BM, Segurens B, Daubin V, Anthouard V, Aiach N *et al*: **Global trends of whole-genome duplications revealed by the ciliate *Paramecium tetraurelia***. *Nature* 2006, **444**(7116):171-178.

81. Arnaiz O, Cain S, Cohen J, Sperling L: **ParameciumDB: a community resource that integrates the *Paramecium tetraurelia* genome sequence with genetic data**. *Nucleic Acids Res* 2007, **35**(suppl_1):D439-444.

82. Eisen JA, Coyne RS, Wu M, Wu DY, Thiagarajan M, Wortman JR, Badger JH, Ren QH, Amedeo P, Jones KM *et al*: **Macronuclear genome sequence of the ciliate *Tetrahymena thermophila*, a model eukaryote**. *PLoS Biol* 2006, **4**(9):1620-1642.

Table 4.1. List of species name abbreviation and data sources. The annotated protein sequences for each genome were downloaded from the respective data source with the version date as indicated. Two ciliates, *Paramecium tetraurelia* and *Tetrahymena thermophila*, are included as outgroups.

Abbr.	Species name	Number of sequences	Version date	Data source
Ch	*Cryptosporidium hominis* [30]	3,886	12/15/2006	CryptoDB [76]
Cp	*Cryptosporidium parvum* [29]	3,806	04/02/2006	CryptoDB [76]
Pf	*Plasmodium falciparum* [28]	5,411	12/07/2005	PlasmoDB [64]
Pv	*Plasmodium vivax*	5,352	12/07/2005	PlasmoDB [64]
Ta	*Theileria annulata* [31]	3,795	07/15/2005	GeneDB [77]
Tp	*Theileria parva* [32]	4,079	08/30/2005	J. Craig Venter Institute [78]
Tg	*Toxoplasma gondii*	7,793	01/04/2006	ToxoDB [79]
Pt	*Paramecium tetraurelia* [80]	39,642	12/11/2006	ParameciumDB [81]
Tt	*Tetrahymena thermophila* [82]	27,769	04/14/2006	J. Craig Venter Institute [78]

Table 4.2. Nucleotide substitution rates in *Theileria*. Genes that are not single-copy or have a high level of divergence (i.e., $d_S > 1$) are excluded because the substitution rate estimates are not reliable. Level 6 genes are not included because they are species-specific and have no orthologous sequence for comparison. The nonsynonymous substitution rate (d_N) indicates the number of nonsynonymous substitutions per nonsynonymous site; the synonymous substitution rate (d_S) indicates the number of synonymous substitutions per synonymous site.

Gene set	Number of sequences		d_N		d_S		d_N/d_S ratio	
	Included	Excluded	Mean	Std. Dev.	Mean	Std. Dev.	Mean	Std. Dev.
Ta1	518	299	0.05	0.04	0.69	0.15	0.07	0.05
Ta2	159	83	0.06	0.04	0.70	0.15	0.09	0.05
Ta3	227	119	0.08	0.04	0.71	0.14	0.11	0.06
Ta4	107	68	0.09	0.05	0.71	0.15	0.13	0.06
Ta5	687	593	0.13	0.07	0.73	0.14	0.19	0.10

Table 4.3. Relative codon bias. The relative codon bias between groups of genes was calculated based on the method developed by Karlin *et al.* [44]. The gene sets listed in columns are used as the reference and the gene sets listed in rows are the focal set. A. Relative codon bias between sets of *P. falciparum* genes. B. Relative codon bias between sets of *T. annulata* genes.

A.

	All	Pf1	Pf2	Pf3	Pf4	Pf5	Pf6
All	*	0.037	0.026	0.015	0.019	0.015	0.087
Pf1	0.040	*	0.030	0.033	0.045	0.047	0.115
Pf2	0.026	0.028	*	0.017	0.037	0.032	0.107
Pf3	0.015	0.031	0.017	*	0.027	0.020	0.102
Pf4	0.019	0.042	0.037	0.026	*	0.021	0.094
Pf5	0.014	0.042	0.031	0.019	0.020	*	0.095
Pf6	0.091	0.115	0.110	0.104	0.101	0.103	*

B.

	All	Ta1	Ta2	Ta3	Ta4	Ta5	Ta6
All	*	0.055	0.084	0.032	0.034	0.043	0.159
Ta1	0.055	*	0.037	0.040	0.077	0.098	0.206
Ta2	0.084	0.037	*	0.064	0.105	0.127	0.231
Ta3	0.032	0.040	0.064	*	0.046	0.068	0.187
Ta4	0.034	0.078	0.106	0.047	*	0.040	0.162
Ta5	0.043	0.098	0.127	0.068	0.040	*	0.130
Ta6	0.165	0.215	0.239	0.196	0.167	0.135	*

Table 4.4. Characteristics of lineage-specific genes in *Plasmodium falciparum*. Gene sets from Pf1 through Pf6 refer to the orthologous gene clusters present in the six levels of lineage specificity defined in Figure 4.2. Pf6A is the same as Pf6 except that it excludes three surface antigen gene families (i.e., PfEMP1, rifin, and stevor). Note that there may be more than one *P. falciparum* gene in a gene cluster when paralogous genes are present in the genome.

Gene set	Number of gene clusters	Number of *P. falciparum* genes	Average protein length (a.a.)	Frequency of genes with			
				"hypothetical protein" in product description	Pfam domains	expression data	predicted signal peptide or transmembrane domains
Pf1	768	803	718	0.26	0.96	0.92	0.16
Pf2	239	244	998	0.70	0.84	0.91	0.29
Pf3	340	346	650	0.66	0.74	0.88	0.49
Pf4	172	175	803	0.74	0.65	0.93	0.39
Pf5	1645	1687	839	0.88	0.53	0.87	0.41
Pf6	454	886	481	0.63	0.46	0.72	0.62
Pf6A	451	616	340	0.91	0.25	0.71	0.56

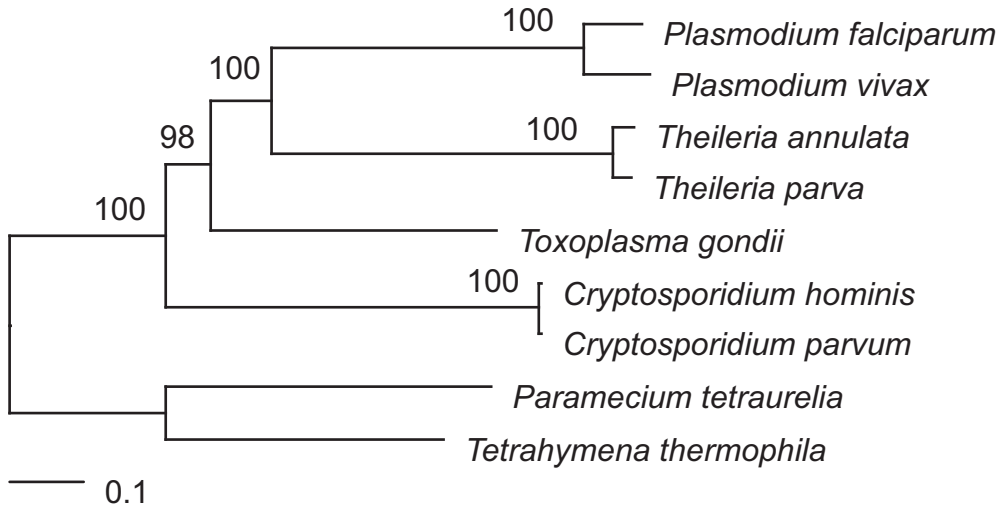


Figure 4.1. The apicomplexan species tree. Maximum likelihood tree generated from the concatenated alignment of 83 single-copy genes (24,494 aligned amino acid sites). Two free-living ciliates, *Paramecium tetraurelia* and *Tetrahymena thermophila*, are included as the outgroup to root the tree. Labels above branches indicate the level of clade support inferred by 100 bootstrap replicates.

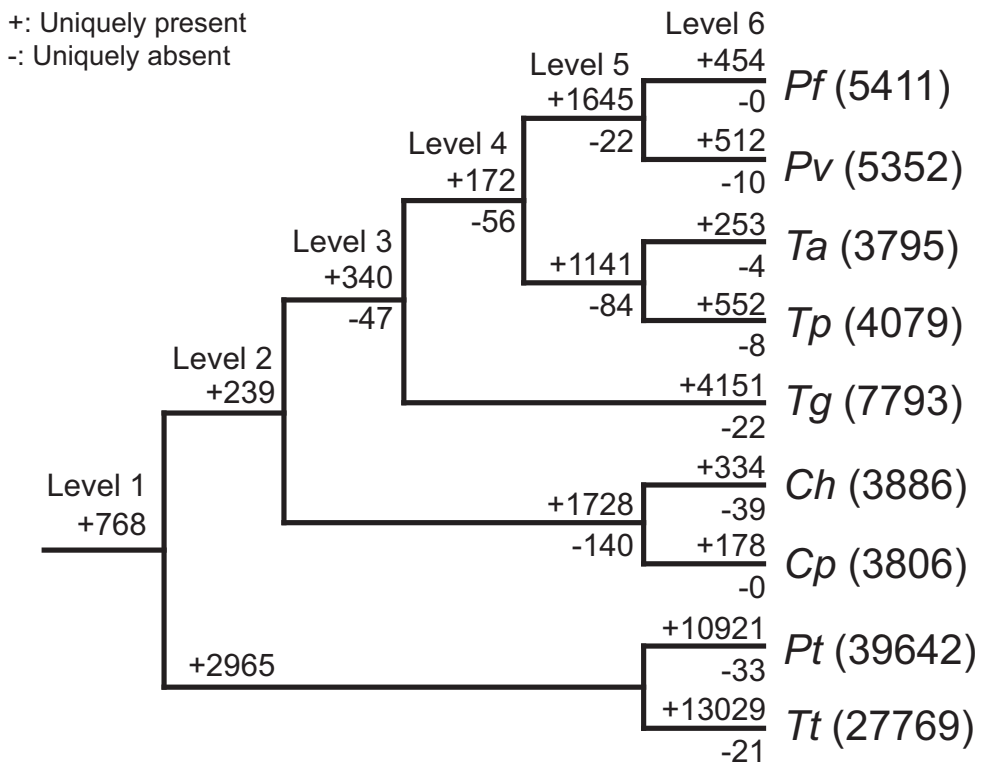


Figure 4.2. Phylogenetic distribution of orthologous gene clusters. The numbers after species name abbreviation (see Table 4.1) indicate the total number of annotated protein coding genes in the genome. The numbers above a branch and proceeded by a '+' sign indicate the number of orthologous gene clusters that are uniquely present in all daughter lineages; the numbers below a branch and proceeded by a '-' sign indicate the number of orthologous gene clusters that are uniquely absent. For example, on the internal branch that leads to the two *Plasmodium* species, 1,645 gene clusters contain sequences from both *Pf* and *Pv* but not any other species present on the tree. Similarly, there are 22 gene clusters that contain sequences from all species except *Pf* and *Pv*. Note that a gene cluster may contain more than one sequence from a species if paralogs are present in the genome. The levels refer to the degree of lineage specificity; genes in level 1 are shared by all species on the tree and genes in level 6 are species-specific.

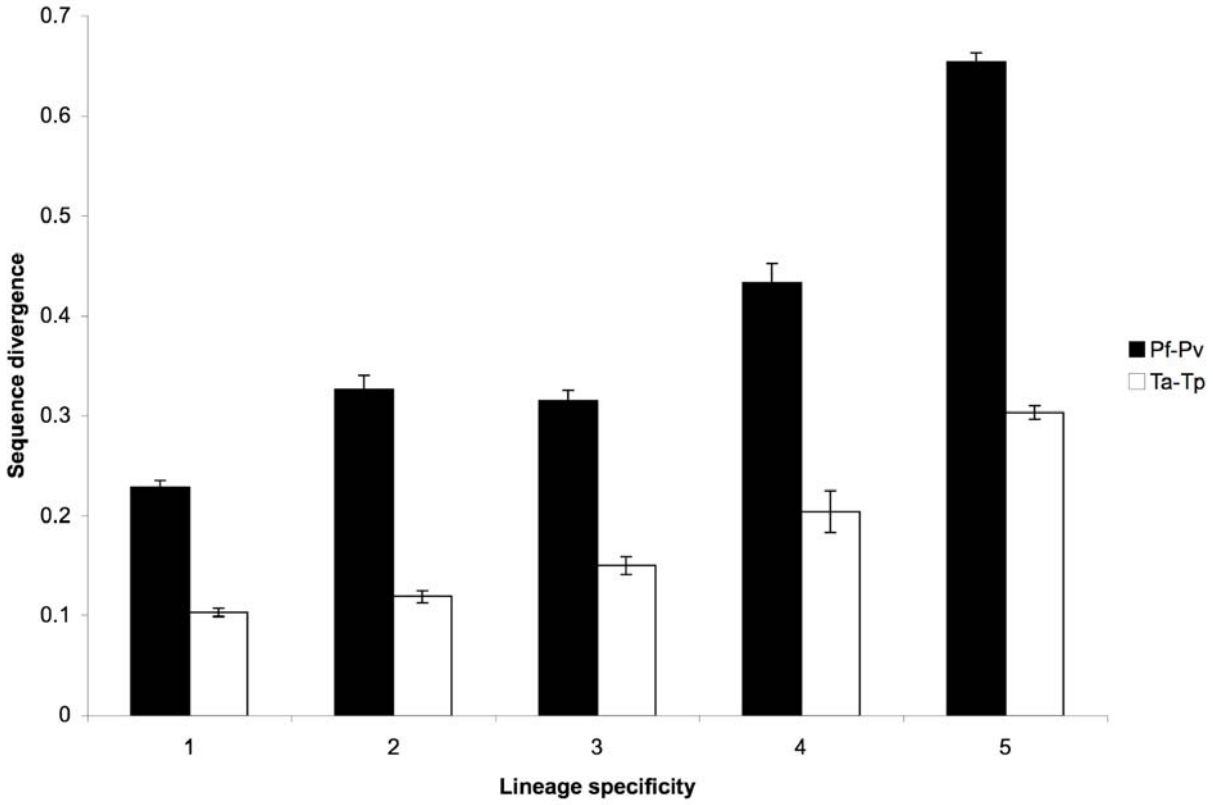


Figure 4.3. Level of amino acid sequence divergence. The five categories on the X-axis refer to the level of lineage specificity defined in Figure 4.2. Level 6 genes are not included because they are species-specific and have no orthologous sequence for comparison. Error bars indicate standard errors.

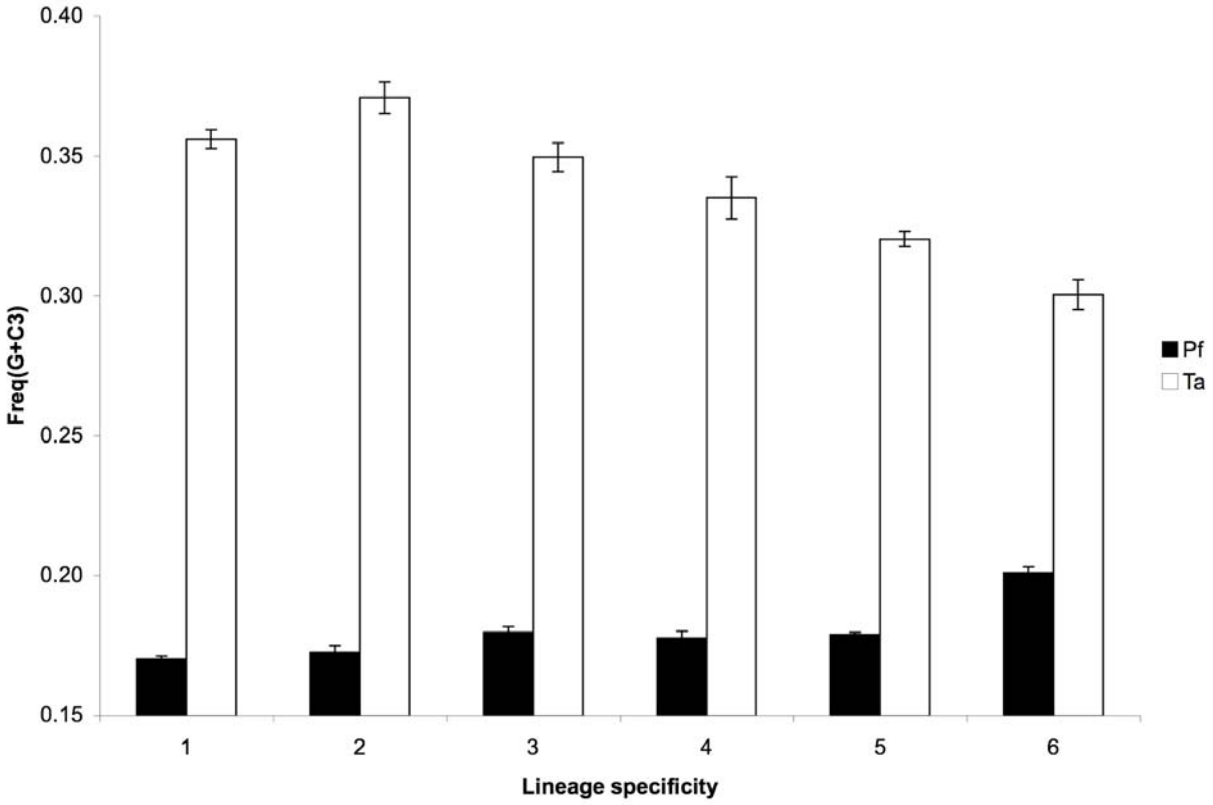


Figure 4.4. (G + C) content at the third codon position. The level of lineage specificity for each calculation is as defined in Figure 4.2. Error bars indicate standard errors.

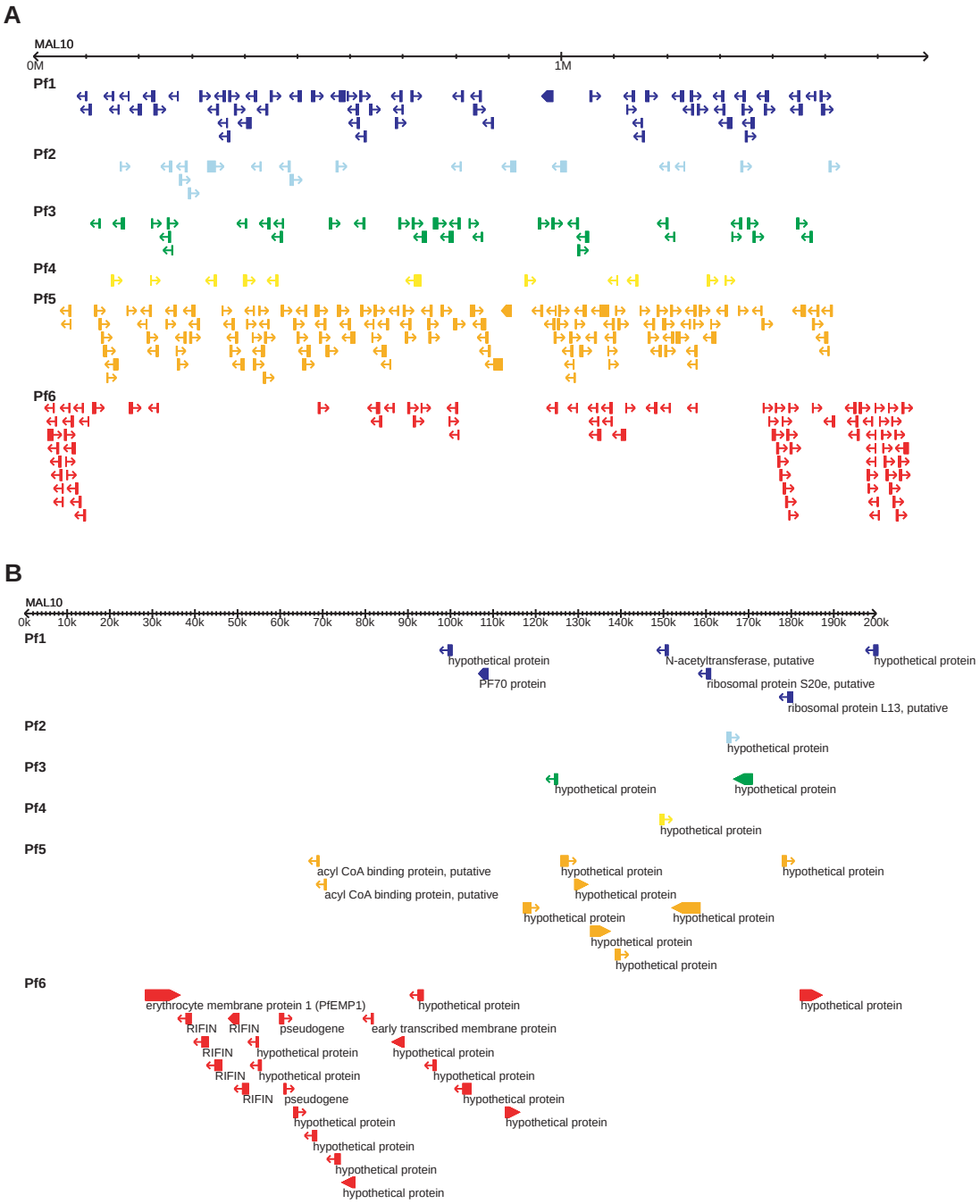


Figure 4.5. Chromosomal location of genes in *Plasmodium falciparum*. Chromosomal location of genes on *P. falciparum* chromosome 10. See Appendix E for views of all 14 chromosomes in this species. The level of lineage specificity is as defined in Figure 4.2. A. View of entire chromosome 10 (MAL10). B. Close-up view of the first 200 kb of chromosome 10.

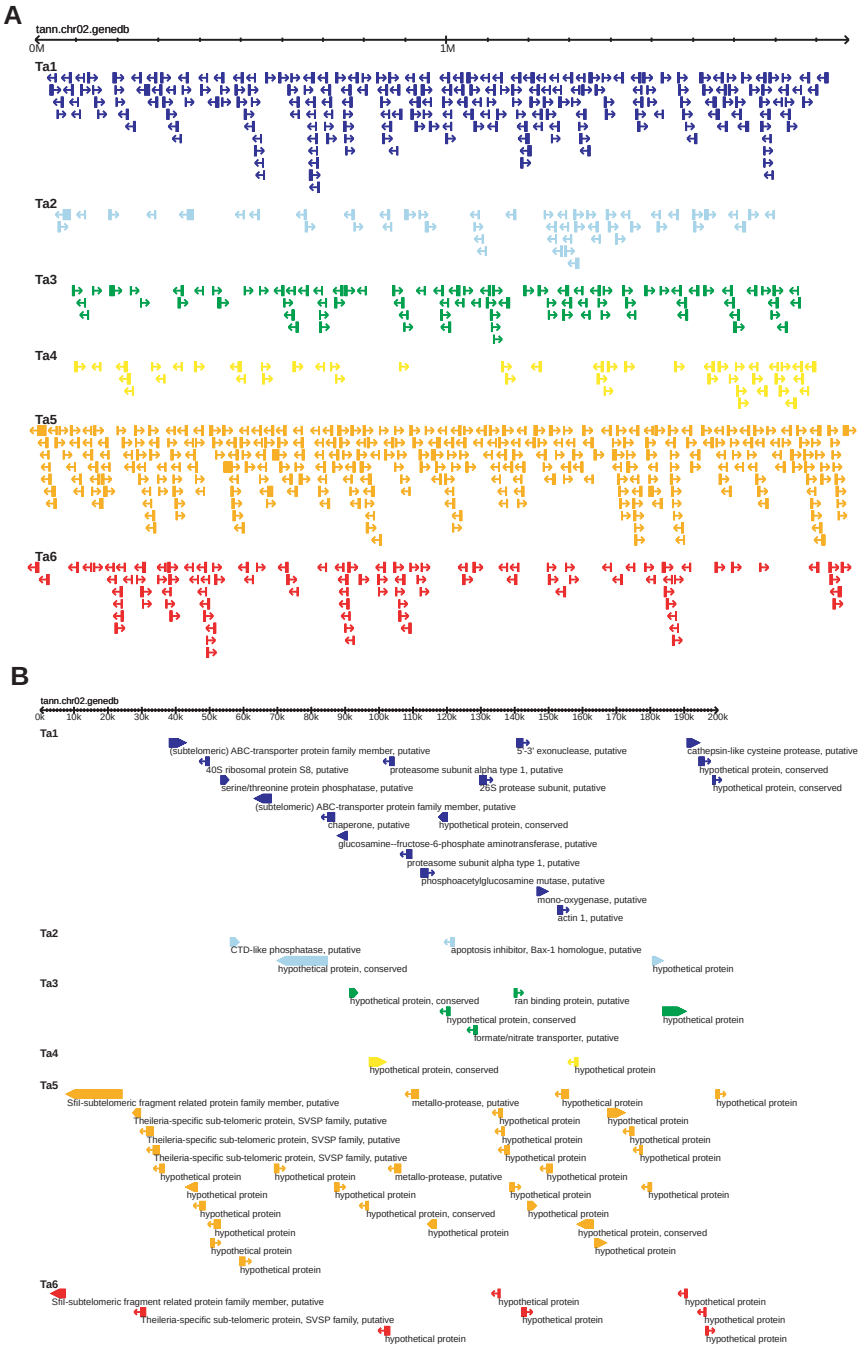


Figure 4.6. Chromosomal location of genes in *Theileria annulata*. Chromosomal location of genes on *T.annulata* chromosome 2. See Appendix F for views of all four chromosomes in this species. The level of lineage specificity is as defined in Figure 4.2. A. View of entire chromosome 2. B. Close-up view of the first 200 kb of chromosome 2.

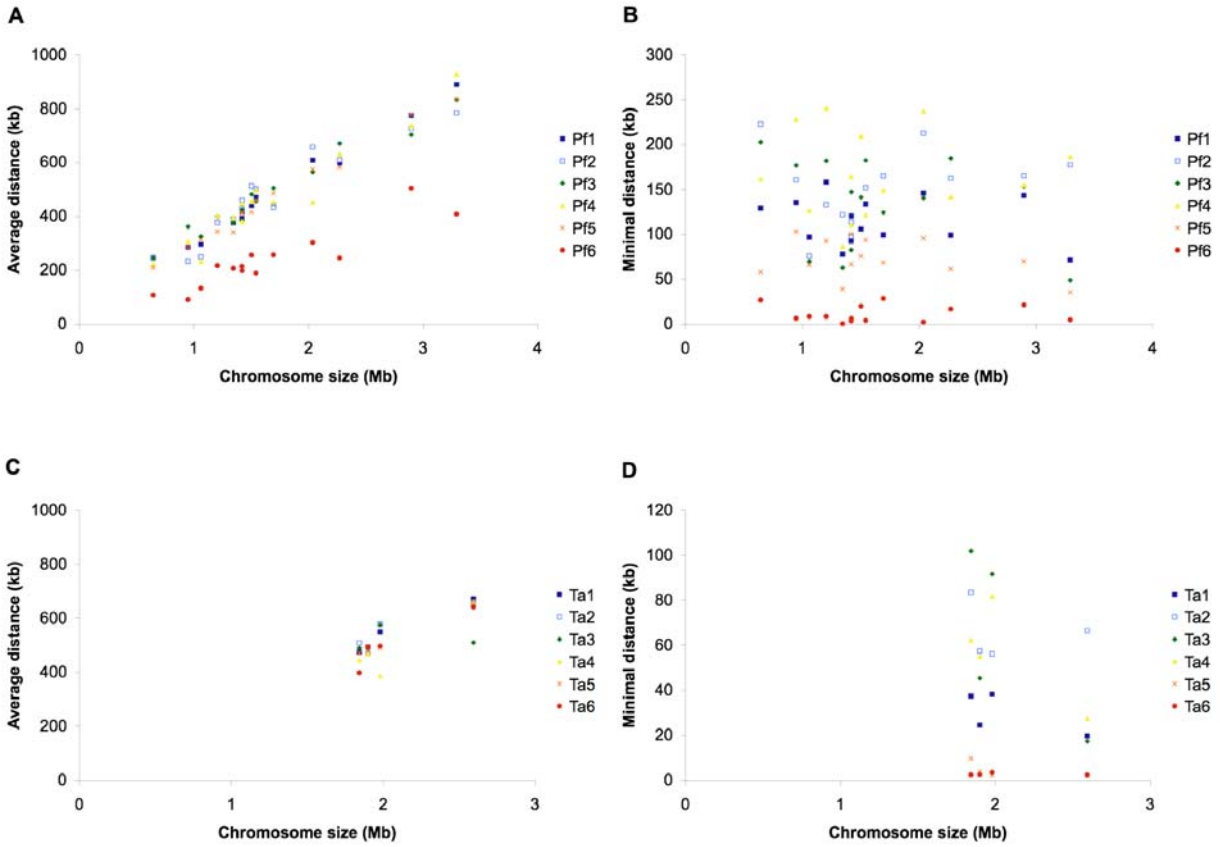


Figure 4.7. Average and minimal distance of mapped genes to chromosome end. The level of lineage specificity is as defined in Figure 4.2. A. Average distance to chromosome end in *Plasmodium falciparum*. B. Minimum distance to chromosome end in *P. falciparum*. C. Average distance to chromosome end in *Theileria annulata*. B. Minimum distance to chromosome end in *T. annulata.*

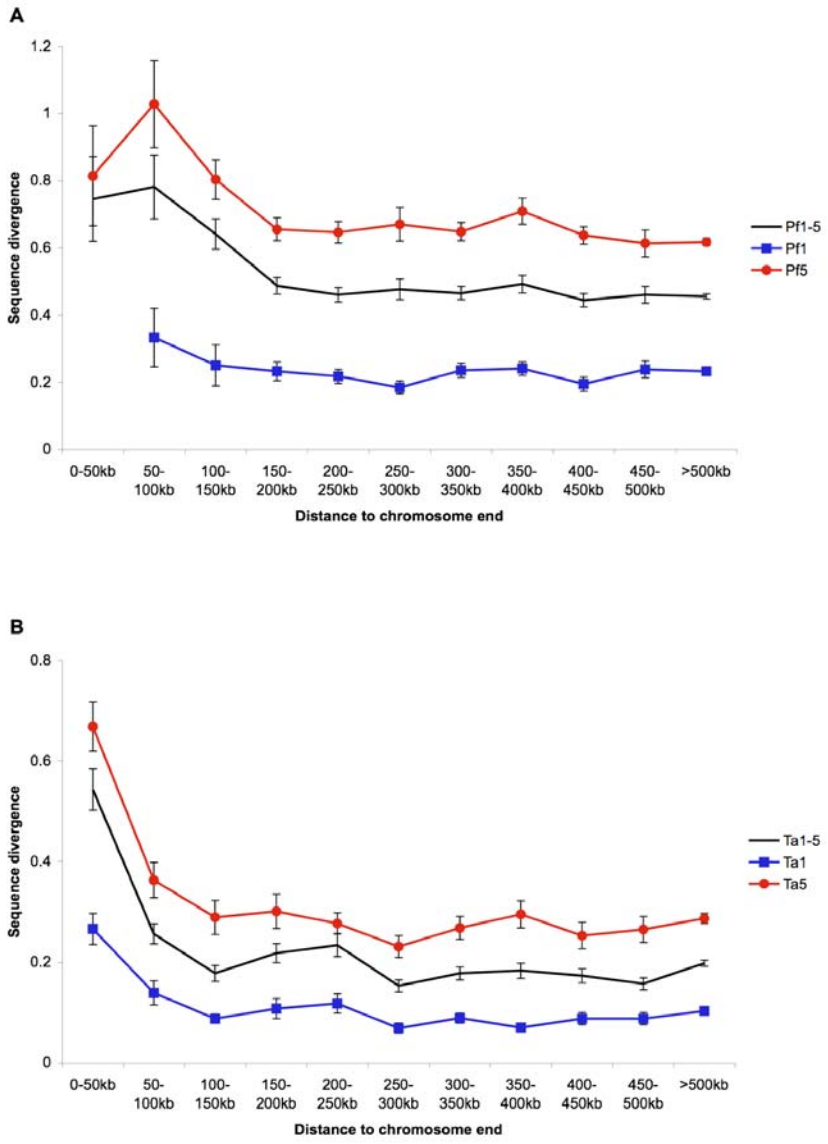


Figure 4.8. Amino acid sequence divergence and chromosomal location. Plot of amino acid sequence divergence as a function of the distance to the nearest chromosome end. A. *Plasmodium falciparum*. B. *Theileria annulata.* The black lines in both panels (i.e., Pf1-5 in panel A and Ta1-5 in panel B) refer to the combined results from genes with five different levels of lineage specificity and are included as the background reference. Error bars indicate standard errors.

CHAPTER 5

CONCLUSIONS

Host-parasite co-evolution is one of the most important topics in evolutionary biology [1, 2] and is directly relevant to improvements of public health [3]. The three studies presented here have examined several facets of this complex system; the implications and possible future directions are discussed below.

Mathematical models are essential tools for developing explicit hypotheses in biology [4]. Results from our theoretical study suggest that parasites may be important in facilitating ecological invasions. Recent studies of one invasive species have provided strong support for this hypothesis [5, 6], which is contrary to the conventional theory that attributes invasion success to the absence of natural enemies [7]. While our simple models have allowed us to draw some rather general conclusions, future modeling efforts that incorporate parasite responses to host evolution and the spatial structures of populations are necessary. Furthermore, empirical studies are essential for testing the hypotheses that have been put forward here and to improve our understanding of the complex evolutionary dynamics of ecological invasion.

Through integration of phylogenetics and genomics, our studies of the apicomplexan parasites have provided several insights into the genome evolution of several devastating pathogens. Our genome-scale phylogenetic analyses indicate that the level of congruence among gene trees is low in apicomplexans compared to that in vertebrates. As a result, a relatively large

number of unlinked loci should be used for phylogenetic inference; molecular phylogenies that were derived from a single locus or a small number of loci should be considered with caution. The list of phylogenetically informative genes that were identified in our study could be good candidates for future studies that aim at improving taxon sampling in this group of important pathogens.

Using the robust species tree inferred from genomic data as the foundation, we have been able to delineate the lineage specificity of genes in two important apicomplexan parasites. Our results of sequence divergence provide further support for the model that gene duplication followed by rapid divergence is a major mechanism for generating novel genes in a genome [8]. Consistent with the hypothesis that lineage-specific genes can be important in adaptation [9], many genus- and species-specific genes in the two parasite genomes analyzed appeared to be surface antigens that the parasites use for interacting with their hosts. These results suggest that refinement of functional annotation in parasite genomes and the collection of genetic variation data will be useful in facilitating our understanding of parasite adaptation and will contribute to the identification of new therapeutic targets.

Due to the complex nature of host-parasite co-evolution, individual studies such as those presented here often focus on one facet at a time for detailed investigations. However, future studies that aim at integrating knowledge from each individual component will be essential to gain a more complete picture of the system.

References

1. May RM, Anderson RM: **Parasite host coevolution**. *Parasitology* 1990, **100**:S89-S101.

2. Anderson RM, May RM: **Coevolution of hosts and parasites**. *Parasitology* 1982, **85**(OCT):411-426.

3. Anderson RM, May RM: **Infectious Diseases of Humans**. Oxford: Oxford University Press; 1991.

4. Otto SP, Day T: **A Biologist's Guide to Mathematical Modeling**. Princeton, N.J.: Princeton University Press; 2007.

5. Tompkins DM, Sainsbury AW, Nettleton P, Buxton D, Gurnell J: **Parapoxvirus causes a deleterious disease in red squirrels associated with UK population declines**. *P Roy Soc London B-Biol Sci* 2002, **269**(1490):529-533.

6. Tompkins DM, White AR, Boots M: **Ecological replacement of native red squirrels by invasive greys driven by disease**. *Ecol Lett* 2003, **6**(3):189-196.

7. Wolfe L: **Why alien invaders succeed: support for the escape-from-enemy hypothesis**. *Am Nat* 2002, **160**(6):705-711.

8. Alba MM, Castresana J: **Inverse relationship between evolutionary rate and age of mammalian genes**. *Mol Biol Evol* 2005, **22**(3):598-606.

9. Wilson GA, Bertrand N, Patel Y, Hughes JB, Feil EJ, Field D: **Orphans as taxonomically restricted and ecologically important genes**. *Microbiology* 2005, **151**(8):2499-2501.

APPENDIX A

EVOLUTIONARILY STABLE STRATEGY ANALYSIS

To investigate the evolutionarily stable strategy of resource allocation (i.e., x_{ESS}) at the population level, we expand our basic model described in Chapter 2 to include two host genotypes. The resident genotype is denoted by the subscript 'R' and the mutant is denoted by the subscript 'M'. We assume that the two host genotypes have the same resource acquisition rate (r) and mortality rate (μ). Parasites can cross-infect the two host genotypes with the same transmission rate (β) and exhibit identical virulence (m) in each host. In other words, the two host genotypes only differ in their resource allocation strategy (i.e., $x_R \neq x_M$). We begin the model with the assumption that the population is at equilibrium with only resident hosts and introduce the mutant at a very low density. The model can be expressed as the following set of differential equations:

$$\frac{dS_R}{dt} = rx_R(1 - S_R - S_M - I_R - I_M)S_R - \beta S_R(I_R + I_M) + r(1 - x_R)I_R - \mu S_R \quad \text{(A1)}$$

$$\frac{dI_R}{dt} = \beta S_R(I_R + I_M) - r(1 - x_R)I_R - m\mu I_R \quad \text{(A2)}$$

$$\frac{dS_M}{dt} = rx_M(1 - S_R - S_M - I_R - I_M)S_M - \beta S_M(I_R + I_M) + r(1 - x_M)I_M - \mu S_M \quad \text{(A3)}$$

$$\frac{dI_M}{dt} = \beta S_M(I_R + I_M) - r(1 - x_M)I_M - m\mu I_M \quad \text{(A4)}$$

To determine if the mutant genotype can invade the resident population, we linearize the system at equilibrium in the absence of mutant (i.e., $S_R = S^*$ and $I_R = I^*$ in eq[6] in Chapter 2, $S_M = I_M = 0$). This results in an upper triangular Jacobian matrix (**J**):

$$\mathbf{J} = \begin{bmatrix} \mathbf{R} & \mathbf{A} \\ \mathbf{0} & \mathbf{M} \end{bmatrix} \tag{A5}$$

where **0** is a 2 × 2 matrix of zeros and **R**, **A** and **M** are defined by the following:

$$\mathbf{R} = \begin{bmatrix} rx_R - 2rx_R S_R - rx_R I_R - \beta I_R - \mu & -rx_R S_R - \beta S_R + r - rx_R \\ \beta I_R & \beta S_R - r + rx_R - m\mu \end{bmatrix} \tag{A6}$$

$$\mathbf{A} = \begin{bmatrix} -rx_R S_R & -rx_R S_R - \beta S_R \\ 0 & \beta S_R \end{bmatrix} \tag{A7}$$

$$\mathbf{M} = \begin{bmatrix} rx_M - rx_M S_R - rx_M I_R - \beta I_R - \mu & r - rx_M \\ \beta I_R & -r + rx_M - m\mu \end{bmatrix} \tag{A8}$$

The eigenvalues of the Jacobian maxtrix are those of submatrices **R** and **M**. Because the resident reaches equilibrium in the absence of the mutant, both eigenvalues of submatrix **R** have negative real parts. Thus, the invasion success of the mutant is determined by the eigenvalues of submatrix **M**. If both eigenvalues of submatrix **M** have negative real parts then the system is stable in the absence of the mutant. In other words, x_R is an ESS when both eigenvalues of submatrix **M** have negative real parts for any x_M given that $x_R \neq x_M$. Based on this definition, we obtained an analytical solution of the ESS for host strategy (i.e., x_{ESS}) by solving for the x_R that satisfies the condition

$$\frac{\partial \lambda}{\partial x_R} = 0 \tag{A9}$$

where λ is the dominant eigenvalue of submatrix **M**. The resulting solution is given in eq (9) of the main text in Chapter 2.

APPENDIX B

LIST OF PHYLOGENETICALLY INFORMATIVE GENES IN THE SUBPHYLUM VERTEBRATA

Topology	Gene ID	Contig	Start	End	Annotation
Fig. 3.3A	ENSG00000187488	1	1250006	1254140	MGC10334 protein.
Fig. 3.3A	ENSG00000060688	1	31505004	31542204	WD repeat protein 57
Fig. 3.3A	ENSG00000162813	1	218297449	218329807	3'(2'),5'-bisphosphate nucleotidase 1
Fig. 3.3A	ENSG00000059588	1	232593682	232681472	TAR RNA binding protein 1
Fig. 3.3A	ENSG00000197769	1	240225415	240228998	Microtubule-associated proteins 1A/1B light chain 3C precursor
Fig. 3.3A	ENSG00000152133	2	37165098	37179891	Coiled-coil domain containing 75
Fig. 3.3A	ENSG00000115946	2	68238509	68256593	Putatative 28 kda protein
Fig. 3.3A	ENSG00000169599	2	69476403	69518257	HIRA-interacting protein 5.
Fig. 3.3A	ENSG00000115561	2	86584064	86644111	Charged multivesicular

					body protein 3
Fig. 3.3A	ENSG00000115866	2	136380724	136459692	Aspartyl-trna synthetase
Fig. 3.3A	ENSG00000138382	2	170376513	170389670	Methyltransferase like 5
Fig. 3.3A	ENSG00000128656	2	175372338	175578361	N-chimaerin
Fig. 3.3A	ENSG00000128694	2	190319631	190335738	O-sialoglycoprotein endopeptidase-like 1
Fig. 3.3A	ENSG00000127837	2	218837097	218843137	Angio-associated migratory cell protein.
Fig. 3.3A	ENSG00000170876	3	14141546	14160178	Transmembrane protein 43
Fig. 3.3A	ENSG00000172113	3	48310595	48317852	Nucleoside diphosphate kinase 6
Fig. 3.3A	ENSG00000178537	3	48869370	48911333	Mitochondrial carnitine/acylcarnitine carrier protein
Fig. 3.3A	ENSG00000132394	3	129355003	129610178	Selenocysteine-specific elongation factor
Fig. 3.3A	ENSG00000138231	3	139362547	139376463	Debranching enzyme homolog 1
Fig. 3.3A	ENSG00000082996	3	151013194	151162613	RING finger protein 13.
Fig. 3.3A	ENSG00000174007	3	197918318	197923491	None
Fig. 3.3A	ENSG00000168228	4	24923504	24980764	Zinc finger CCHC domain-containing

					protein 4.
Fig. 3.3A	ENSG00000151466	4	130024837	130234212	Sodium channel associated protein 1
Fig. 3.3A	ENSG00000137460	4	154083585	154120288	None
Fig. 3.3A	ENSG00000145604	5	36187946	36219902	S-phase kinase-associated protein 2
Fig. 3.3A	ENSG00000112972	5	43325255	43349241	Hydroxymethylglutaryl-coa synthase, cytoplasmic
Fig. 3.3A	ENSG00000145740	5	68425839	68462648	Zinc transporter ZTL1 isoform 1
Fig. 3.3A	ENSG00000094880	5	137551259	137576918	Cell division cycle protein 23 homolog
Fig. 3.3A	ENSG00000120725	5	138310342	138561904	Nucleotide exchange factor SIL1 precursor
Fig. 3.3A	ENSG00000113312	5	159368708	159425096	Tetratricopeptide repeat protein 1
Fig. 3.3A	ENSG00000106080	7	30019409	30032793	FK506-binding protein 14 precursor
Fig. 3.3A	ENSG00000126524	7	66090125	66098023	Shwachman-Bodian-Diamond syndrome protein.
Fig. 3.3A	ENSG00000106077	7	72788363	72791120	Abhydrolase domain

					containing 11 isoform 1
Fig. 3.3A	ENSG00000146802	7	112193032	112217684	Transmembrane protein 168
Fig. 3.3A	ENSG00000104331	8	58037822	58068957	Myo-inositol monophosphatase A3
Fig. 3.3A	ENSG00000164751	8	78057748	78074907	Peroxisome assembly factor 1
Fig. 3.3A	ENSG00000147647	8	105460829	105548453	Dihydropyrimidinase
Fig. 3.3A	ENSG00000168496	11	61316726	61321284	Flap endonuclease 1
Fig. 3.3A	ENSG00000137522	11	71317731	71386289	Ring finger protein 121 isoform 1
Fig. 3.3A	ENSG00000166159	12	1799956	1816179	Leucine-rich repeats and transmembrane domains 2
Fig. 3.3A	ENSG00000047621	12	4467195	4517898	None
Fig. 3.3A	ENSG00000205074	12	79412848	79414455	None
Fig. 3.3A	ENSG00000139324	12	87060218	87116936	Transmembrane and tetratricopeptide repeat containing 3
Fig. 3.3A	ENSG00000139372	12	102883747	102906785	G/T mismatch-specific thymine DNA glycosylase
Fig. 3.3A	ENSG00000136045	12	106603720	106630387	Periodic tryptophan

					protein 1 homolog
Fig. 3.3A	ENSG00000033030	12	121523390	121551340	Zinc finger CCHC domain-containing protein 8.
Fig. 3.3A	ENSG00000179630	13	43351969	43364367	UPF0124 protein.
Fig. 3.3A	ENSG00000100568	14	67186985	67211096	Vesicle transport through interaction with t-snares homolog 1B
Fig. 3.3A	ENSG00000119698	14	93710402	93815825	HEAT-like repeat-containing protein isoform 1
Fig. 3.3A	ENSG00000166477	15	50017516	50051271	RNA polymerase-associated protein LEO1
Fig. 3.3A	ENSG00000140463	15	70765588	70817869	Bardet-Biedl syndrome 4 protein.
Fig. 3.3A	ENSG00000117906	15	75011174	75029349	Reticulocalbin-2 precursor
Fig. 3.3A	ENSG00000166411	15	76228774	76249938	Isocitrate dehydrogenase
Fig. 3.3A	ENSG00000173867	15	86803713	86811623	39S ribosomal protein L46, mitochondrial precursor
Fig. 3.3A	ENSG00000065057	16	2029817	2037868	Endonuclease III-like protein 1

Fig. 3.3A	ENSG00000182810	16	66612984	66614606	Probable ATP-dependent RNA helicase DDX28
Fig. 3.3A	ENSG00000176208	17	26183146	26246421	Chromosome fragility associated gene 1
Fig. 3.3A	ENSG00000167291	17	75526303	75624242	TBC1 domain family member 16.
Fig. 3.3A	ENSG00000101624	18	12662632	12692703	Centrosomal protein of 76 kda
Fig. 3.3A	ENSG00000101782	18	19287254	19317031	Serine/threonine-protein kinase RIO3
Fig. 3.3A	ENSG00000125912	19	3136875	3160572	Nicalin precursor
Fig. 3.3A	ENSG00000159259	21	36679559	36710994	Chromatin assembly factor 1 subunit B
Fig. 3.3A	ENSG00000070010	22	17817464	17846693	Ubiquitin fusion degradation 1-like isoform A
Fig. 3.3A	ENSG00000100209	22	27468019	27483488	Co-chaperone protein hscb, mitochondrial precursor
Fig. 3.3A	ENSG00000100297	22	34126128	34150494	DNA replication licensing factor MCM5
Fig. 3.3B	ENSG00000049656	5	1370983	1397999	Cisplatin resistance related protein CRR9p

Fig. 3.3B	ENSG00000065268	19	935328	945569	WD repeat protein 18.
Fig. 3.3C	ENSG00000149397	11	118460797	118469469	Porphobilinogen deaminase
Fig. 3.3C	ENSG00000063854	16	1799109	1817163	Hydroxyacylglutathione hydrolase
Fig. 3.3D	ENSG00000149084	11	43658835	43834736	Estradiol 17-beta-dehydrogenase 12
Fig. 3.3E	ENSG00000128563	7	101791064	101854125	PRKR interacting protein 1
Fig. 3.3F	ENSG00000161326	17	32924064	32947701	Dual specificity protein phosphatase 14
Fig. 3.3G	ENSG00000168092	11	116528374	116543625	Platelet-activating factor acetylhydrolase IB subunit beta

APPENDIX C

LIST OF PHYLOGENETICALLY INFORMATIVE GENES IN THE PHYLUM APICOMPLEXA

Topology	Gene ID	Contig	Start	End	Annotation
Fig. 3.5A	PFC0805w	MAL3	745136	752509	DNA-directed RNA polymerase II, putative
Fig. 3.5A	PFD0450c	MAL4	443114	444878	Pre-mrna splicing factor, putative
Fig. 3.5A	PFD0950w	MAL4	877717	878853	Ran binding protein 1
Fig. 3.5A	PFE0505w	MAL5	437493	439736	Cyclophilin, putative
Fig. 3.5A	PFF0100w	MAL6	92503	96665	Putative ATP dependent RNA helicase
Fig. 3.5A	PFF0940c	MAL6	815619	818244	Cell division cycle protein 48 homologue, putative
Fig. 3.5A	PF07_0033	MAL7	451091	453712	Cg4 protein
Fig. 3.5A	PF08_0036	MAL8	998127	1000601	Transport protein
Fig. 3.5A	PFI1090w	MAL9	903352	904560	S-adenosylmethionine synthetase, putative
Fig. 3.5A	PF11_0227	MAL11	832803	834380	Hypothetical protein
Fig. 3.5A	PF11_0331	MAL11	1245223	1247145	T-complex protein 1, alpha

					subunit, putative
Fig. 3.5A	PF11_0445	MAL11	1728809	1729810	DNA-directed RNA polymerase I, putative
Fig. 3.5A	PFL0580w	MAL12	521681	524464	DNA replication licensing factor mcm5, putative
Fig. 3.5A	PFL0625c	MAL12	558040	562173	Eukaryotic translation initiation factor 3 subunit 10, putative
Fig. 3.5A	PF13_0178	MAL13	1348977	1349720	Translation initiation factor 6, putative
Fig. 3.5A	PF13_0224	MAL13	1633891	1634561	60S ribosomal subunit protein L18, putative
Fig. 3.5A	PF13_0328	MAL13	2478805	2479629	Proliferating cell nuclear antigen
Fig. 3.5A	PF14_0064	MAL14	244907	245491	Vacuolar protein sorting 29, putative
Fig. 3.5A	PF14_0177	MAL14	753387	756438	DNA replication licensing factor MCM2
Fig. 3.5A	PF14_0296	MAL14	1245983	1246903	Ribosomal protein L14, putative
Fig. 3.5B	PF07_0073	MAL7	820172	821791	Seryl-trna synthetase, putative
Fig. 3.5B	PF07_0121	MAL7	1304215	1306881	Hypothetical protein, conserved
Fig. 3.5B	PFL0830w	MAL12	678595	681954	Hypothetical protein
Fig. 3.5B	PFL1425w	MAL12	1211777	1213552	T-complex protein 1, gamma subunit, putative
Fig. 3.5B	MAL13P1.218	MAL13	1721935	1723844	UDP-N-acetylglucosamine

					pyrophosphorylase, putative
Fig. 3.5B	MAL13P1.270	MAL13	2150495	2151550	Proteasome subunit, putative
Fig. 3.5B	PF14_0635	MAL14	2728411	2729761	Hypothetical protein, conserved
Fig. 3.5C	PFB0635w	MAL2	566320	567894	T-complex protein 1, putative
Fig. 3.5C	PFI0155c	MAL9	133761	135421	Ras family GTP-ase, putative
Fig. 3.5C	PFI1260c	MAL9	1028651	1030000	Histone deacetylase
Fig. 3.5C	PF11_0483	MAL11	1885547	1889507	Farnesyltransferase beta subunit, putative
Fig. 3.5C	PF14_0143	MAL14	578386	587527	Hypothetical protein
Fig. 3.5D	PFA0400c	MAL1	329584	330457	Beta3 proteasome subunit, putative
Fig. 3.5D	PF08_0130	MAL8	168189	171554	Wd repeat protein, putative
Fig. 3.5D	PF10_0077	MAL10	316648	318279	Eukaryotic translation initiation factor 3 subunit 7, putative
Fig. 3.5D	PF13_0331	MAL13	2489476	2490045	Hypothetical protein, conserved
Fig. 3.5D	PF14_0393	MAL14	1689948	1691468	Structure specific recognition protein, putative
Fig. 3.5E	PFF1500c	MAL6	1291038	1292975	DEAD/DEAH box ATP-dependent RNA helicase, putative
Fig. 3.5E	PF08_0126	MAL8	213372	217091	DNA repair protein rad54, putative
Fig. 3.5E	PF11_0377	MAL11	1430051	1432080	Casein kinase 1
Fig. 3.5E	PF14_0587	MAL14	2496995	2499332	Hypothetical protein

Fig. 3.5F	PFC0185w	MAL3	201218	202795	Hypothetical protein, conserved
Fig. 3.5F	PFL0670c	MAL12	590187	592427	Bi-functional aminoacyl-trna synthetase, putative
Fig. 3.5F	PFL2215w	MAL12	1920770	1921900	Actin
Fig. 3.5F	PF13_0156	MAL13	1184409	1185221	Proteasome subunit beta type 7 precursor, putative
Fig. 3.5G	PFF1070c	MAL6	903738	906320	Hypothetical protein, conserved
Fig. 3.5G	PF08_0121	MAL8	279497	280150	Peptidyl-prolyl cis-trans isomerase precursor
Fig. 3.5G	PFI1545c	MAL9	1273072	1273920	Proteasome precursor, putative
Fig. 3.5G	PF14_0174	MAL14	726977	728377	Hypothetical protein, conserved
Fig. 3.5H	PFD1090c	MAL4	1053710	1054334	Clathrin assembly protein, putative
Fig. 3.5H	PF11_0055	MAL11	194384	195972	Hypothetical protein
Fig. 3.5H	PFL0930w	MAL12	757642	763635	Clathrin heavy chain, putative
Fig. 3.5I	PF08_0085	MAL8	632704	633349	Ubiquitin-conjugating enzyme, putative
Fig. 3.5I	PFI0875w	MAL9	737975	740266	Heat shock protein
Fig. 3.5I	PF13_0102	MAL13	774991	776946	DNAJ-like Sec63 homologue
Fig. 3.5J	PFC0295c	MAL3	309367	310016	40S ribosomal protein S12, putative

APPENDIX D

GENES USED FOR INFERRING THE APICOMPLEXAN SPECIES TREE

Gene ID	Contig	Start	End	Annotation
PFB0130w	MAL2	135523	137139	Polyprenyl synthetase, putative
PFB0370c	MAL2	335679	336581	RNA-binding protein, putative
PFB0860c	MAL2	750048	751736	RNA helicase, putative
PFB0875c	MAL2	763490	764950	Hypothetical protein
PFC0805w	MAL3	745136	752509	DNA-directed RNA polymerase II, putative
PFC0890w	MAL3	837032	838428	Vesicle transport protein, putative
PFD0455w	MAL4	445784	447232	Ribosomal processing protein, putative
PFD0590c	MAL4	535089	541382	DNA polymerase alpha
PFD0830w	MAL4	755069	756895	Bifunctional dihydrofolate reductase-thymidylate synthase
PFD0950w	MAL4	877717	878853	Ran binding protein 1
PFD1070w	MAL4	1044910	1046082	Eukaryotic initiation factor, putative
PFE0420c	MAL5	347219	355039	Guanidine nucleotide exchange factor, putative
PFE0430w	MAL5	358261	362733	ATP-dependent RNA helicase, putative
PFE0925c	MAL5	763530	766901	Snrnp protein, putative

PFE1050w	MAL5	857035	858474	Adenosylhomocysteinase(S-adenosyl-L-homocystein e hydrolase)
PFE1355c	MAL5	1132932	1134849	Ubiquitin carboxyl-terminal hydrolase, putative
PFF0120w	MAL6	105500	106639	Putative geranylgeranyltransferase
PFF1095w	MAL6	919683	924026	Leucyl-trna synthetase, cytoplasmic, putative
PFF1500c	MAL6	1291038	1292975	DEAD/DEAH box ATP-dependent RNA helicase, putative
MAL7P1.113	MAL7	977663	980362	DEAD box helicase, putative
PF08_0130	MAL8	168189	171554	Wd repeat protein, putative
MAL8P1.125	MAL8	403717	405263	Tyrosyl-trna synthetase, putative
PF08_0098	MAL8	522308	525130	Abc transporter, putative
PFI0290c	MAL9	293582	296614	Beta subunit of coatomer complex, putative
PFI0525w	MAL9	494154	495497	Nucleotide binding protein, putative
PFI0860c	MAL9	729664	732126	ATP-dependant RNA helicase, putative
PFI0865w	MAL9	732941	734017	Hypothetical protein, conserved
PFI0920c	MAL9	772636	774369	Hypothetical protein, conserved
PFI1625c	MAL9	1331339	1332793	Organelle processing peptidase, putative
PF10_0054	MAL10	228798	232193	Hypothetical protein
PF10_0077	MAL10	316648	318279	Eukaryotic translation initiation factor 3 subunit 7, putative
PF10_0099	MAL10	405015	410535	Hypothetical protein
PF10_0150	MAL10	618439	619992	Methionine aminopeptidase, putative

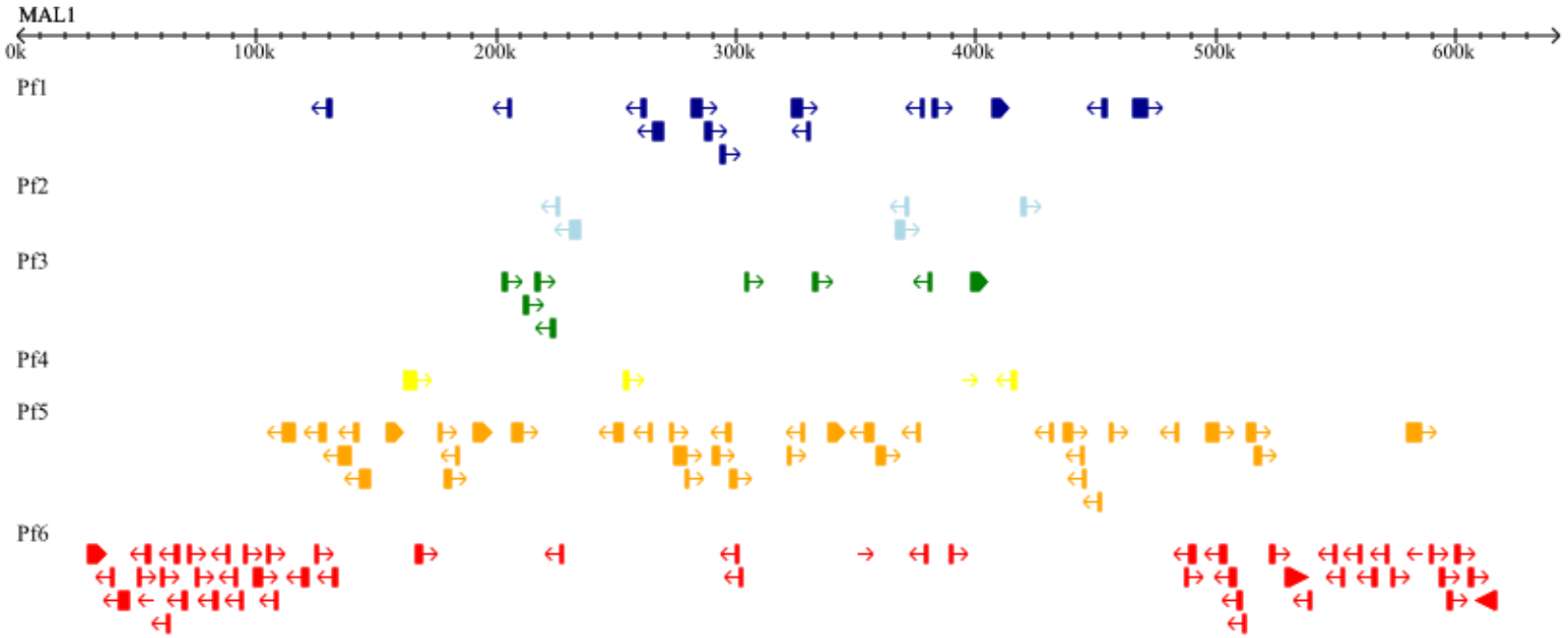
PF10_0165	MAL10	685577	688861	DNA polymerase delta catalytic subunit
PF10_0200	MAL10	835291	839699	Hypothetical protein, conserved
PF10_0209	MAL10	867633	869675	RNA helicase, putative
PF10_0266	MAL10	1135080	1137398	Hypothetical protein
PF11_0108	MAL11	403110	407332	Hypothetical protein, conserved
PF11_0156	MAL11	553128	556282	Hypothetical protein
PF11_0170	MAL11	615926	617392	Cyclophilin, putative
PF11_0184	MAL11	671913	674963	DNA mismatch repair protein MLH1, putative
PF11_0212	MAL11	770153	772656	Hypothetical protein
PF11_0245	MAL11	922642	924438	Translation elongation factor EF-1, subunit alpha, putative
PF11_0258	MAL11	971568	972452	Co-chaperone grpe, putative
PF11_0305	MAL11	1134988	1136892	Hypothetical protein
PF11_0311	MAL11	1156279	1159101	N-acetyl glucosamine phosphate mutase, putative
PF11_0317	MAL11	1177073	1182529	Structural maintenance of chromosome protein, putative
PF11_0336	MAL11	1263163	1264497	Hypothetical protein, conserved
PFL0175c	MAL12	182295	183965	Hypothetical protein, conserved
PFL0330c	MAL12	294994	299346	DNA-directed RNA polymerase III subunit, putative
PFL0355c	MAL12	324332	326854	Hypothetical protein
PFL0720w	MAL12	618757	620104	Hypothetical protein

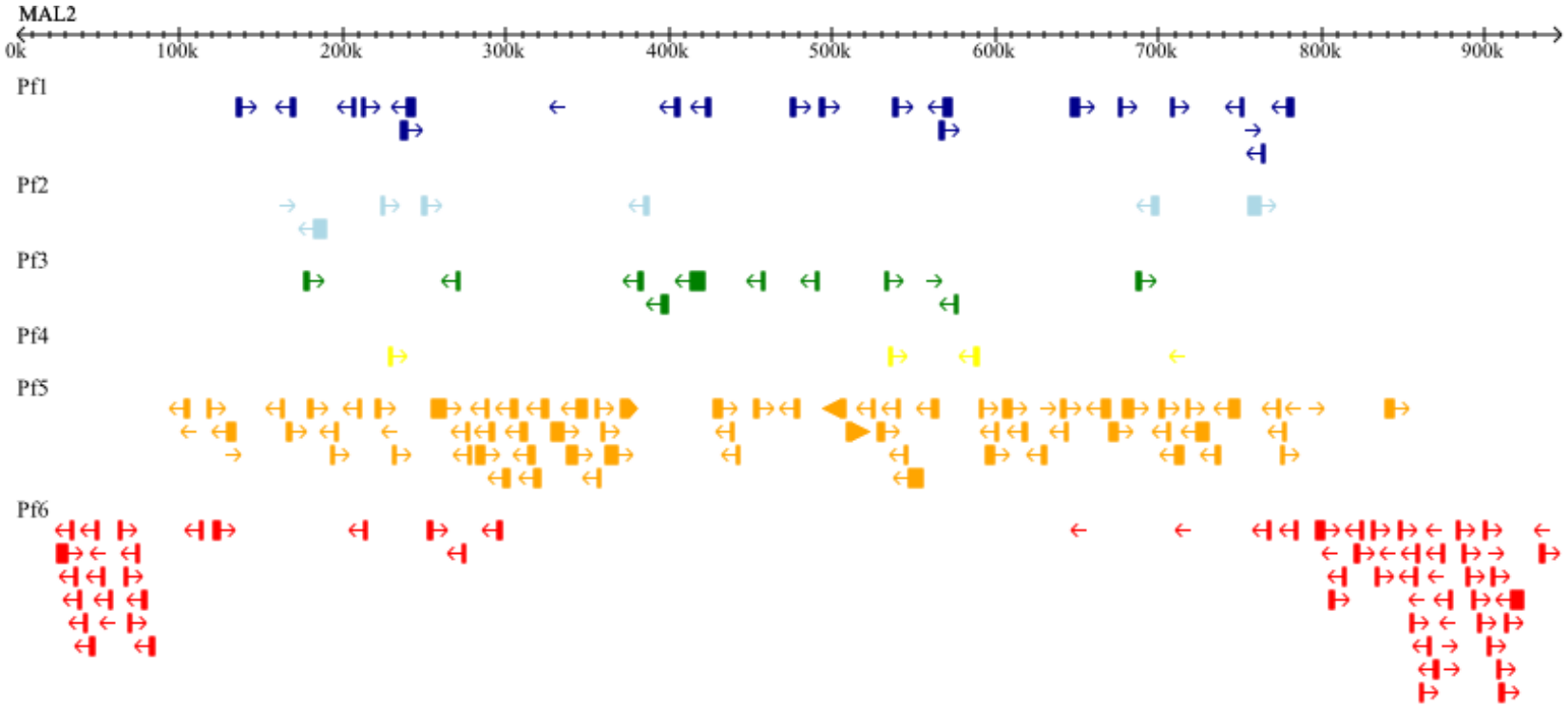
PFL0830w	MAL12	678595	681954	Hypothetical protein
PFL1525c	MAL12	1301088	1304594	Pre-mrna splicing factor RNA helicase, putative
PF13_0013	MAL13	143062	144297	PBS lyase HEAT-like repeat domain protein
MAL13P1.14	MAL13	145376	150400	ATP-dependent DEAD box helicase, putative
PF13_0048	MAL13	408783	412304	Hypothetical protein
MAL13P1.54	MAL13	481701	483284	Hypothetical protein, conserved
PF13_0096	MAL13	708548	710470	Ubiquitin Carboxyl-terminal Hydrolase-like zinc finger protein
MAL13P1.134	MAL13	1023461	1027944	Helicase, putative
MAL13P1.159	MAL13	1272411	1274057	Hypothetical protein, conserved
PF13_0205	MAL13	1488257	1490458	Tryptophan--trna ligase, putative
MAL13P1.191	MAL13	1551144	1553298	Hypothetical protein, conserved
MAL13P1.385	MAL13	1756242	1758185	RNA binding protein, putative
MAL13P1.243	MAL13	1925422	1930052	Elongation factor Tu, putative
PF13_0271	MAL13	2072409	2075558	ABC transporter, putative
MAL13P1.264	MAL13	2077095	2079483	Hypothetical protein
PF13_0309	MAL13	2282553	2286543	Hypothetical protein
MAL13P1.289	MAL13	2350014	2353235	Mitotic control protein dis3 homologue, putative
PF13_0350	MAL13	2659757	2661487	Signal recognition particle receptor alpha subunit, putative
PF14_0052	MAL14	190996	192723	Hypothetical protein, conserved

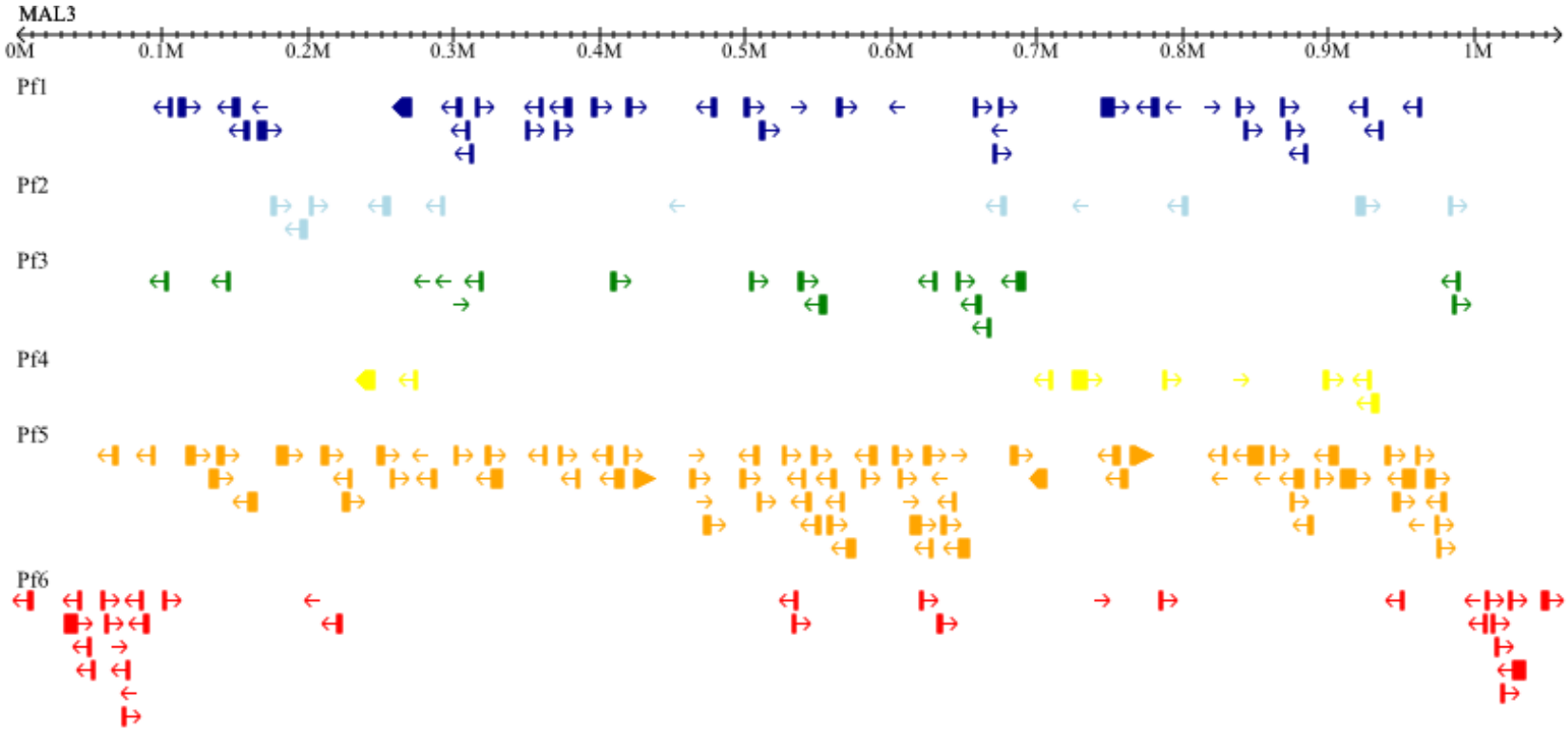
PF14_0100	MAL14	400417	402993	Cytidine triphosphate synthetase
PF14_0115	MAL14	475248	477296	Hypothetical protein
PF14_0143	MAL14	578386	587527	Hypothetical protein
PF14_0148	MAL14	606378	607346	Uracil-DNA glycosylase, putative
PF14_0277	MAL14	1169782	1174246	Coatamer protein, beta subunit, putative
PF14_0341	MAL14	1457086	1458825	Glucose-6-phosphate isomerase
PF14_0364	MAL14	1556482	1559112	Cleavage and polyadenylation specifity factor protein, putative
PF14_0416	MAL14	1793663	1794889	Hypothetical protein
PF14_0428	MAL14	1850610	1854008	Histidine -- trna ligase, putative
PF14_0517	MAL14	2233131	2235425	Peptidase, putative
PF14_0589	MAL14	2513621	2516893	Valine - trna ligase, putative
PF14_0601	MAL14	2569382	2570667	Replication factor C3

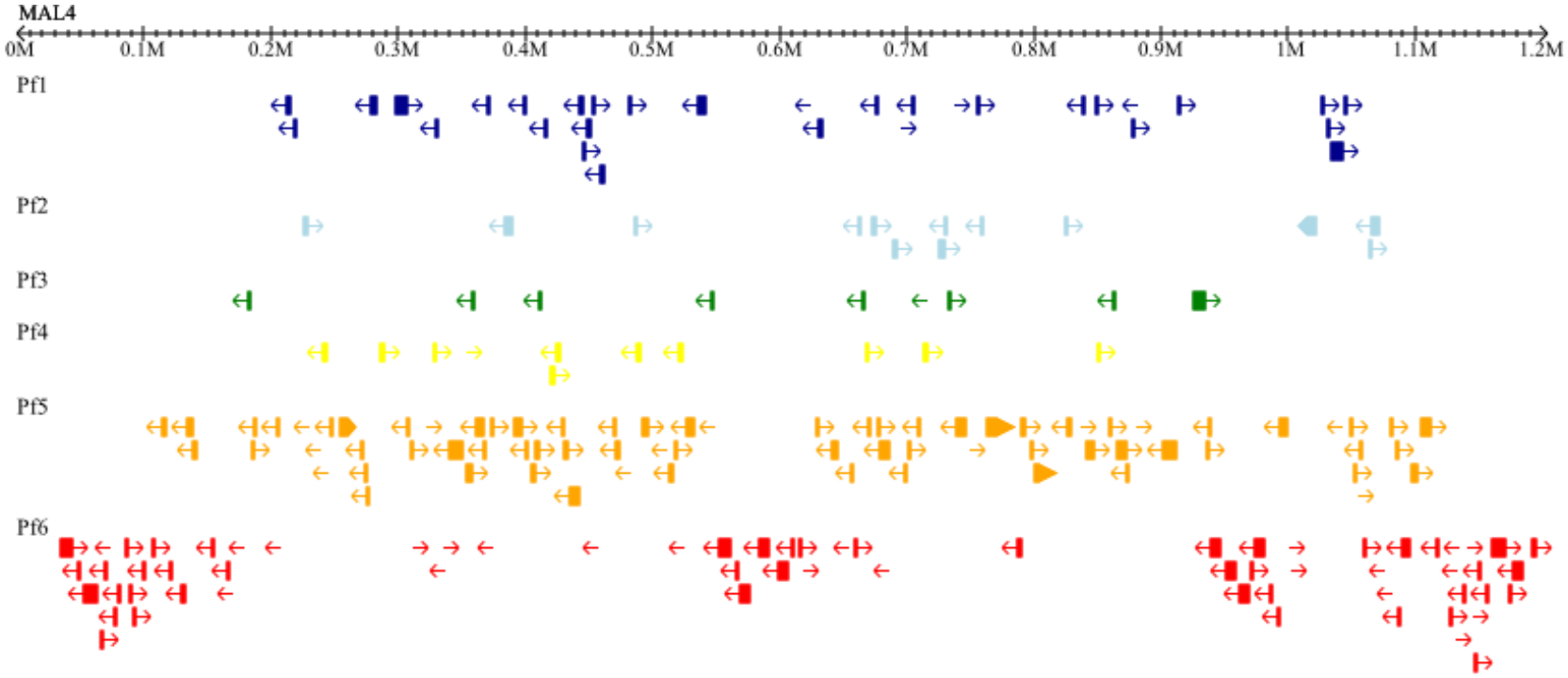
APPENDIX E

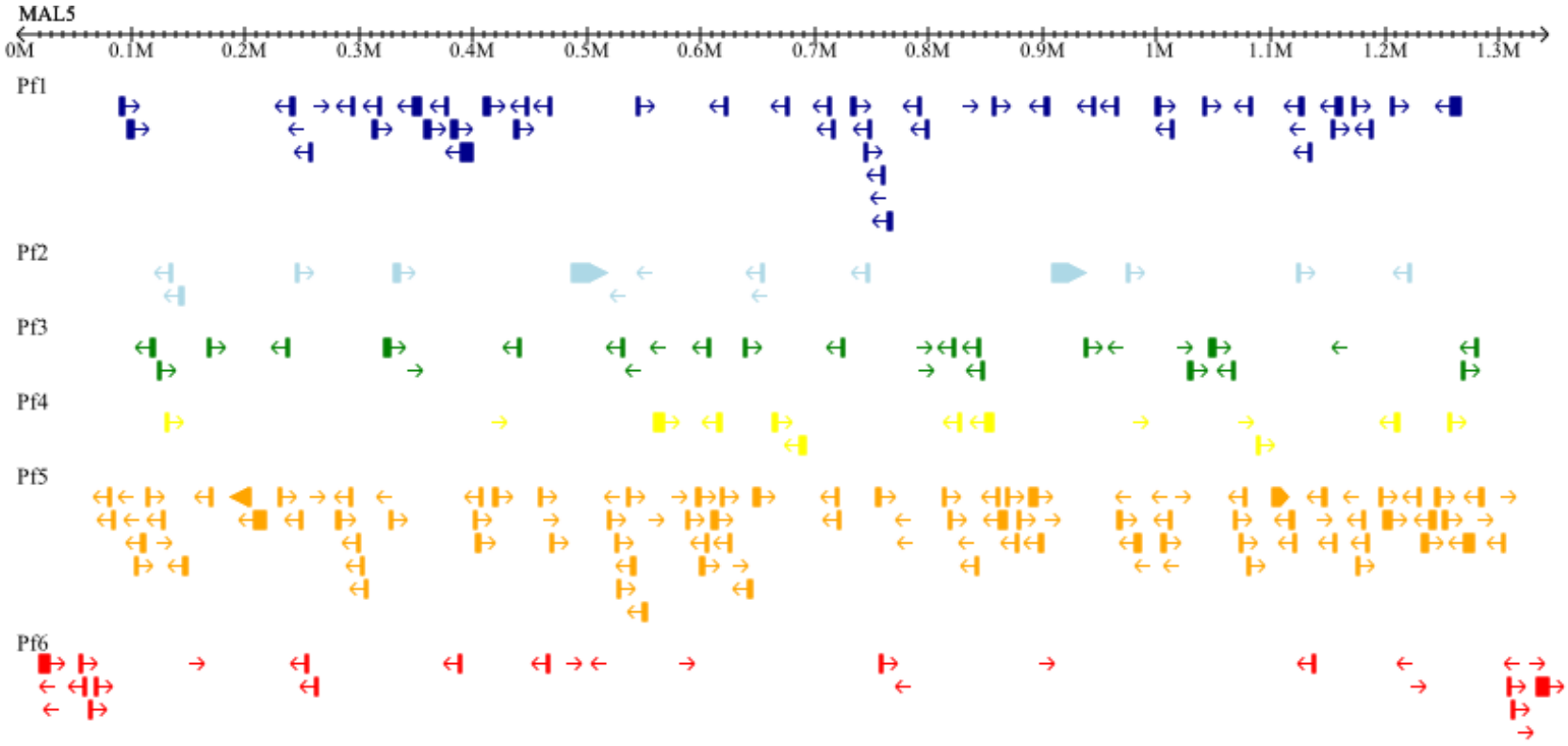
CHROMOSOMAL LOCATION OF LINEAGE-SPECIFIC GENES IN *PLASMODIUM FALCIPARUM*

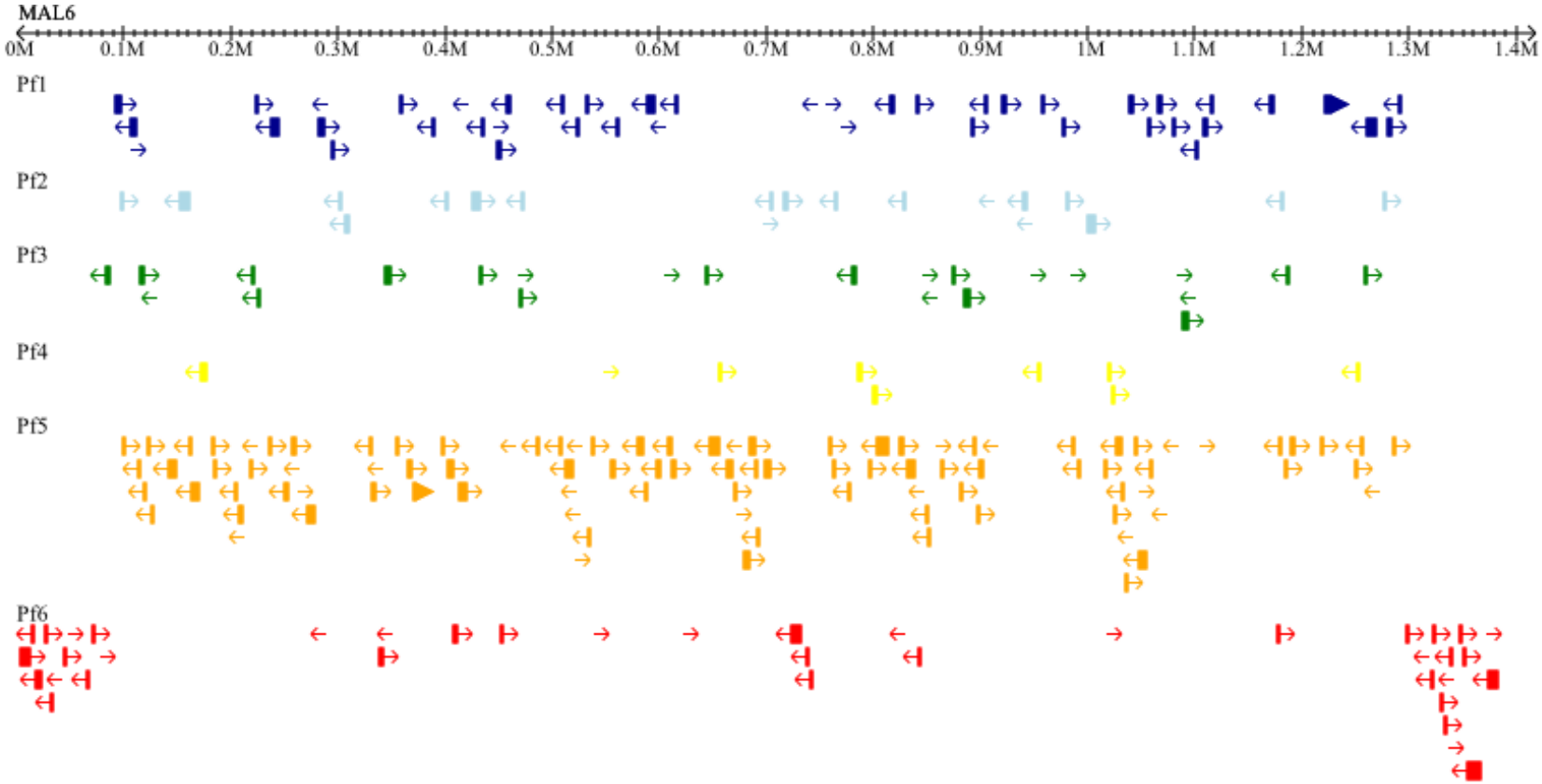


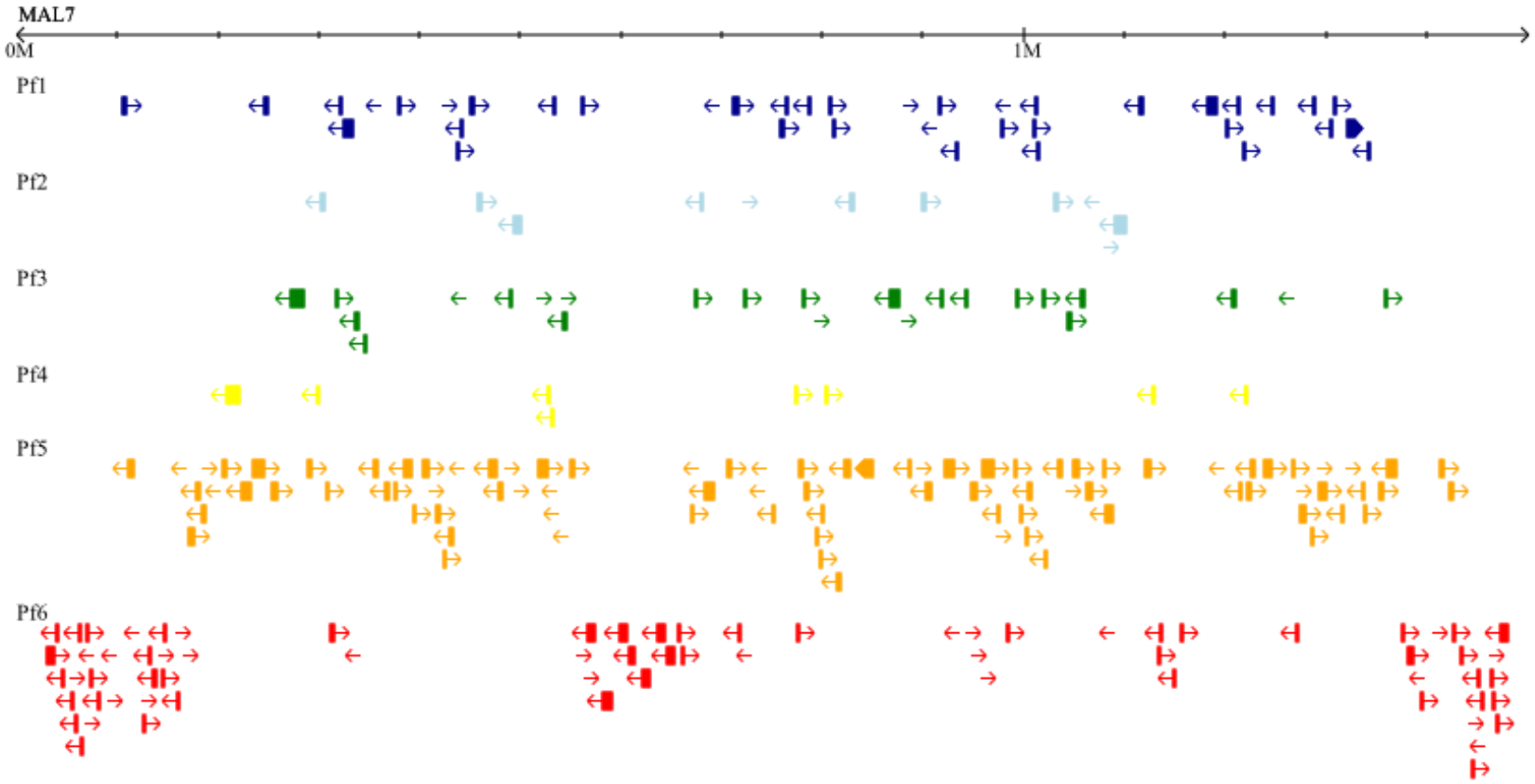

MAL2
0k
100k
200k
300k
400k
500k
600k
700k
800k
900k
Pf1
Pf2
Pf3
Pf4
Pf5
Pf6

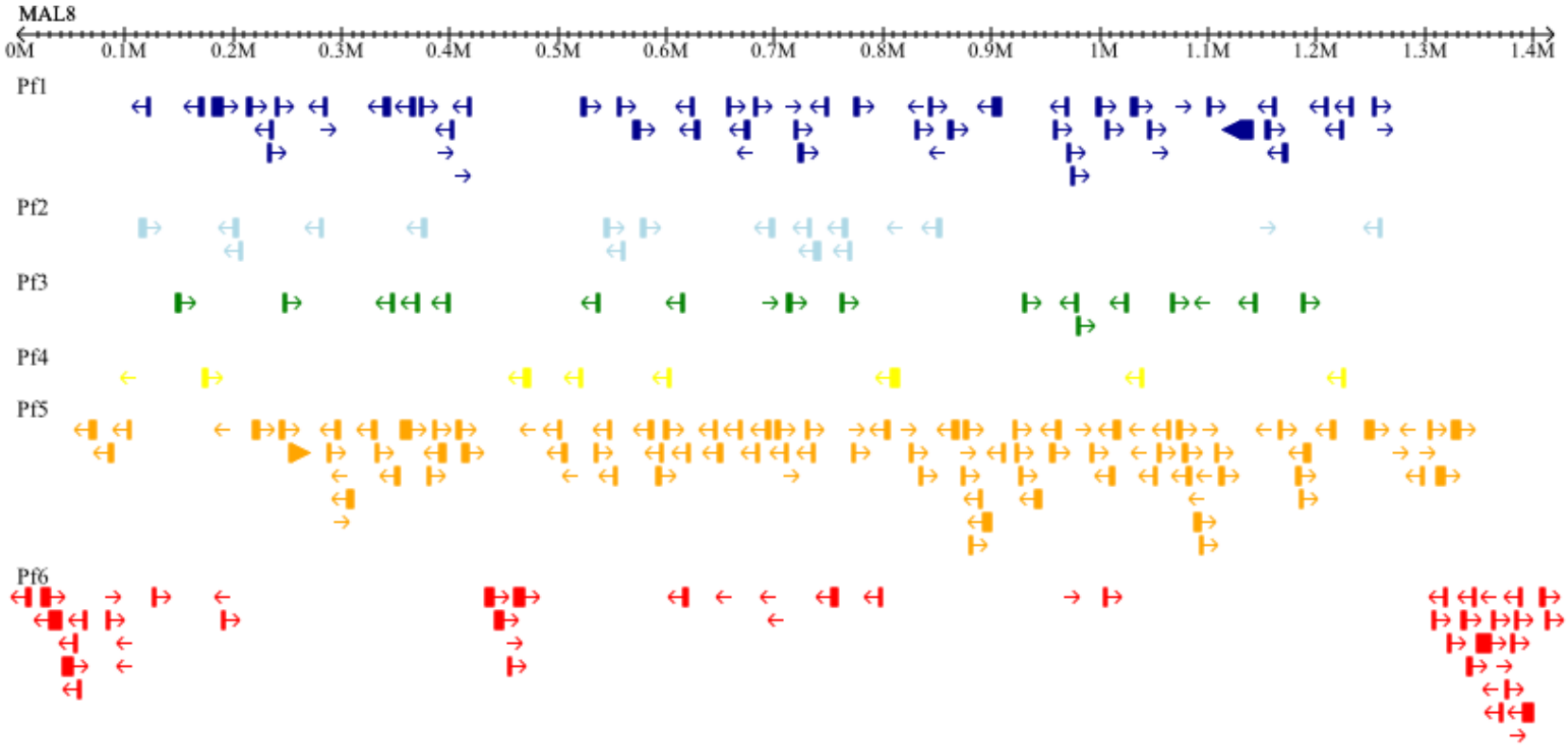

MAL3
0M
0.1M
0.2M
0.3M
0.4M
0.5M
0.6M
0.7M
0.8M
0.9M
1M
Pf1
Pf2
Pf3
Pf4
Pf5
Pf6

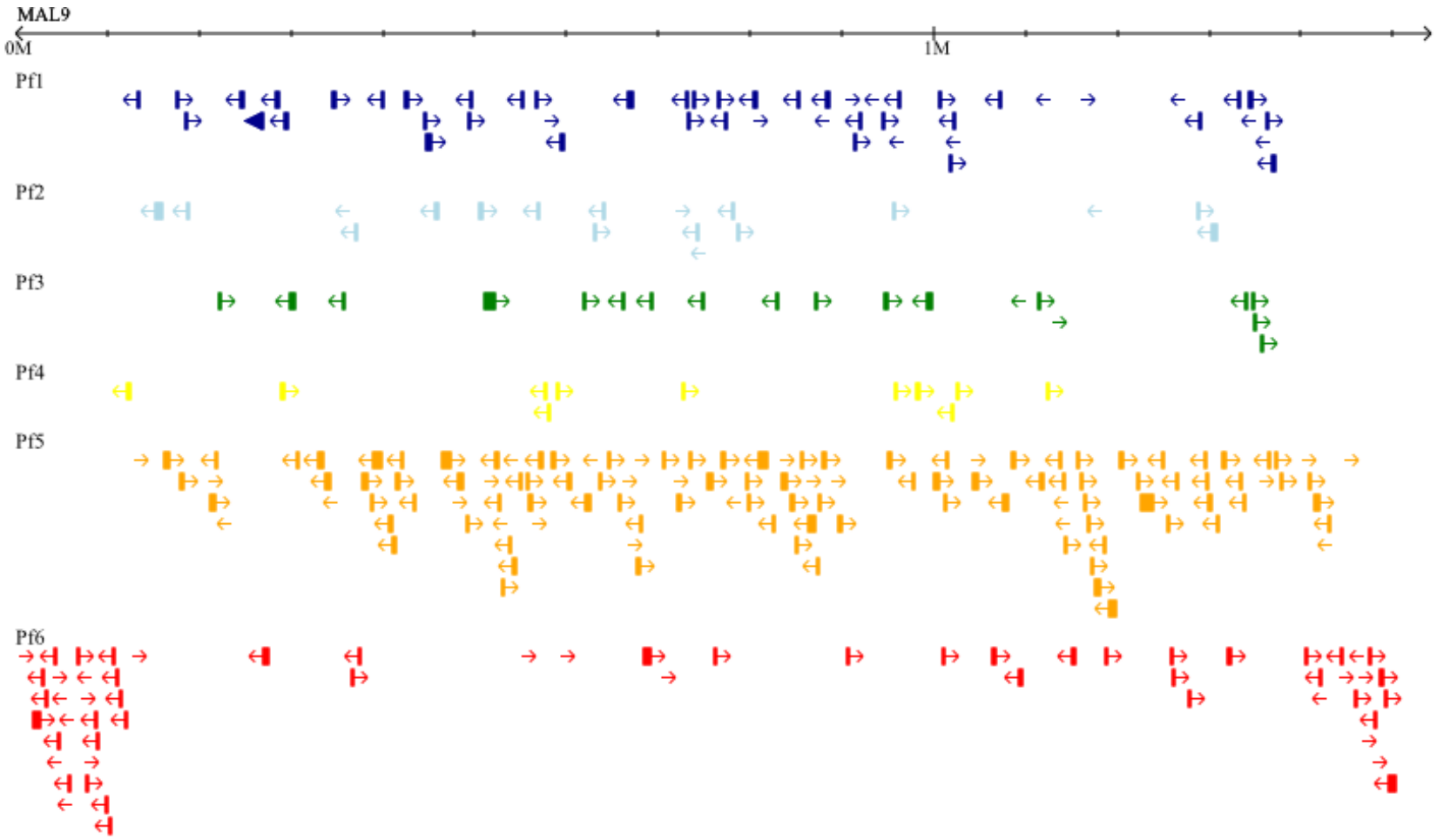

MAL4
0M
0.1M
0.2M
0.3M
0.4M
0.5M
0.6M
0.7M
0.8M
0.9M
1M
1.1M
1.2M
Pf1
Pf2
Pf3
Pf4
Pf5
Pf6

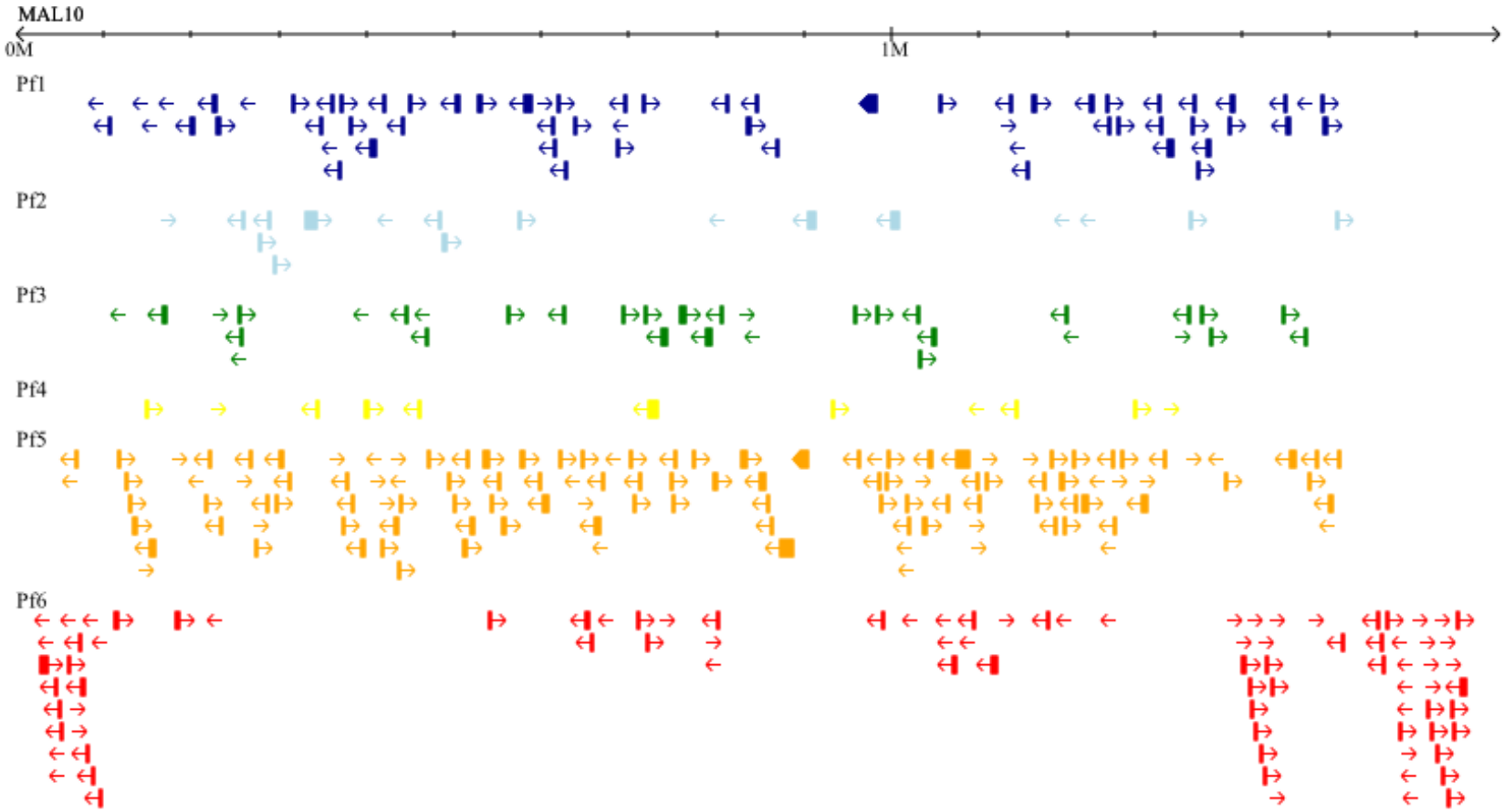

MAL5
0M
0.1M
0.2M
0.3M
0.4M
0.5M
0.6M
0.7M
0.8M
0.9M
1M
1.1M
1.2M
1.3M
Pf1
Pf2
Pf3
Pf4
Pf5
Pf6

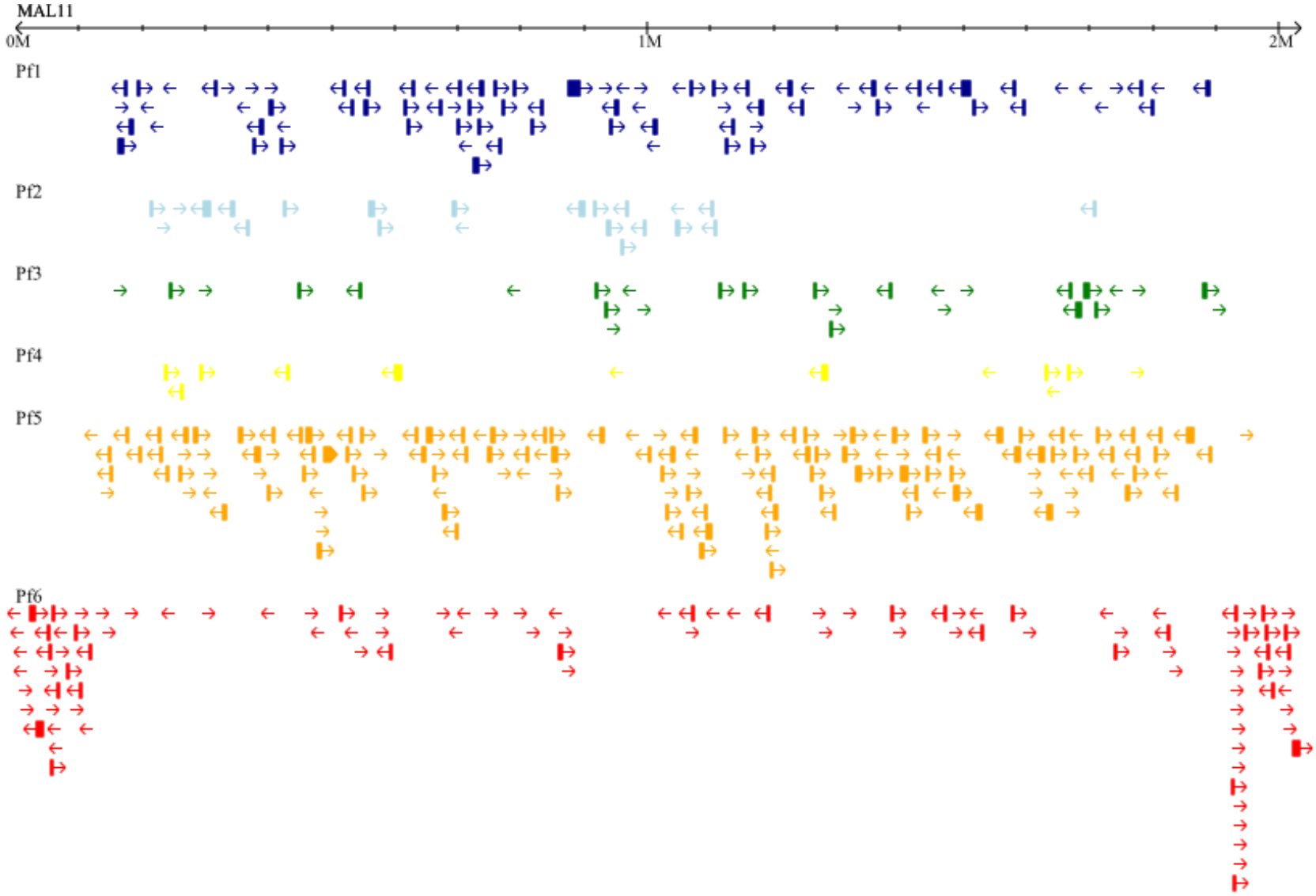

MAL6
0M
0.1M
0.2M
0.3M
0.4M
0.5M
0.6M
0.7M
0.8M
0.9M
1M
1.1M
1.2M
1.3M
1.4M
Pf1
Pf2
Pf3
Pf4
Pf5
Pf6

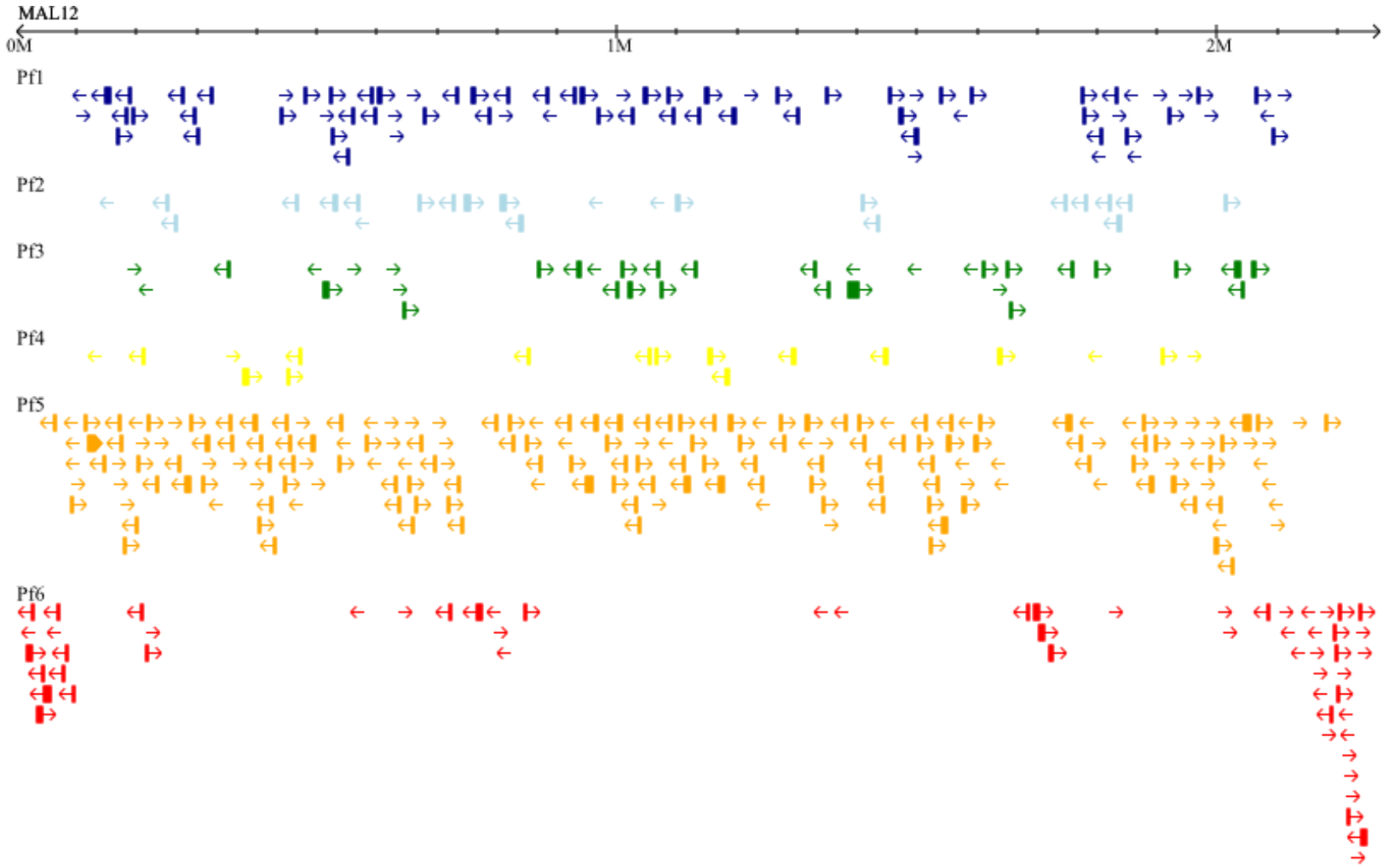

MAL7
0M
1M
Pf1
Pf2
Pf3
Pf4
Pf5
Pf6

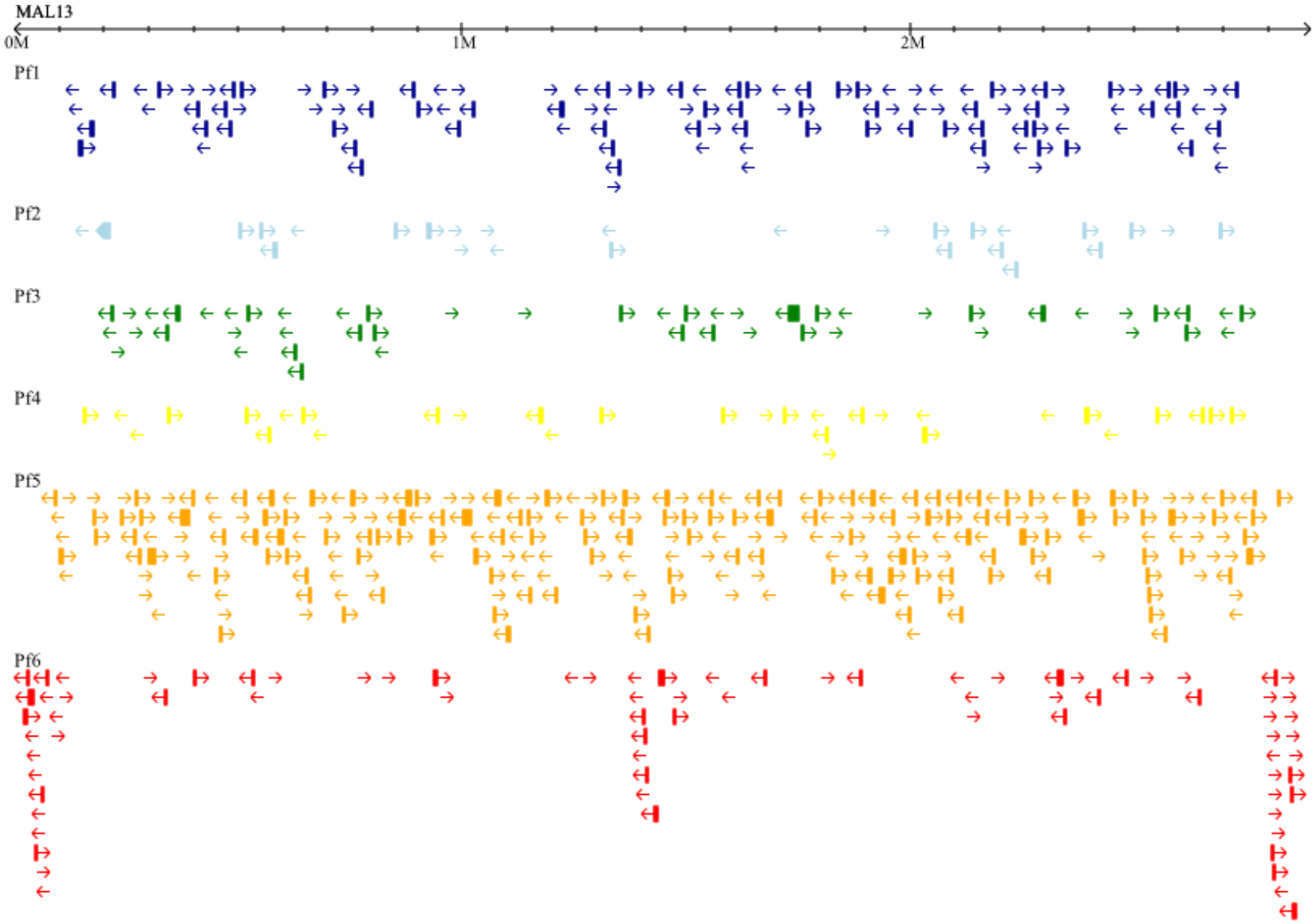

MAL8
0M
0.1M
0.2M
0.3M
0.4M
0.5M
0.6M
0.7M
0.8M
0.9M
1M
1.1M
1.2M
1.3M
1.4M
Pf1
Pf2
Pf3
Pf4
Pf5
Pf6

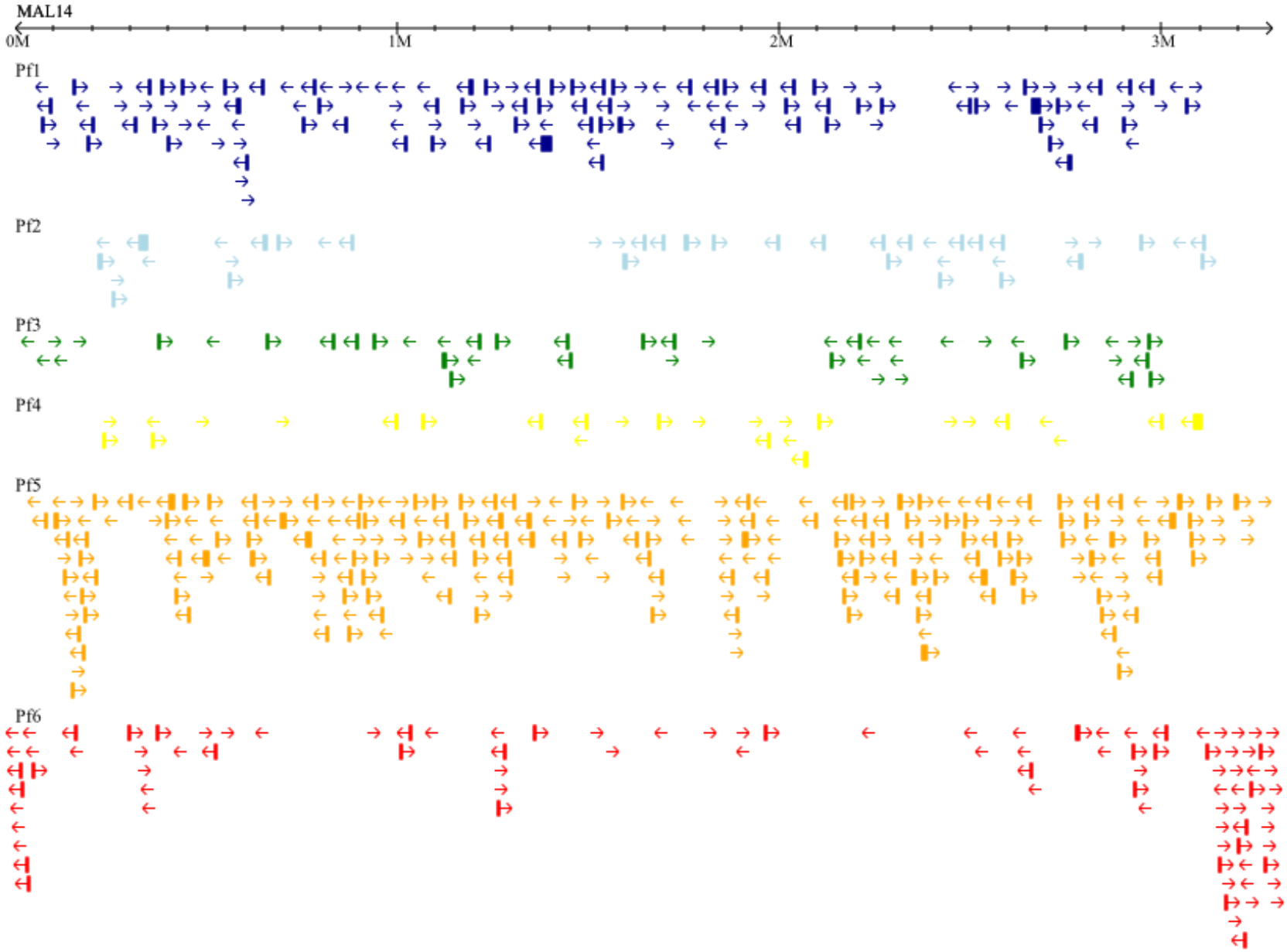

MAL9
0M
1M
Pf1
Pf2
Pf3
Pf4
Pf5
Pf6


MAL10
0M
1M
Pf1
Pf2
Pf3
Pf4
Pf5
Pf6


MAL11
0M
1M
2M
Pf1
Pf2
Pf3
Pf4
Pf5
Pf6

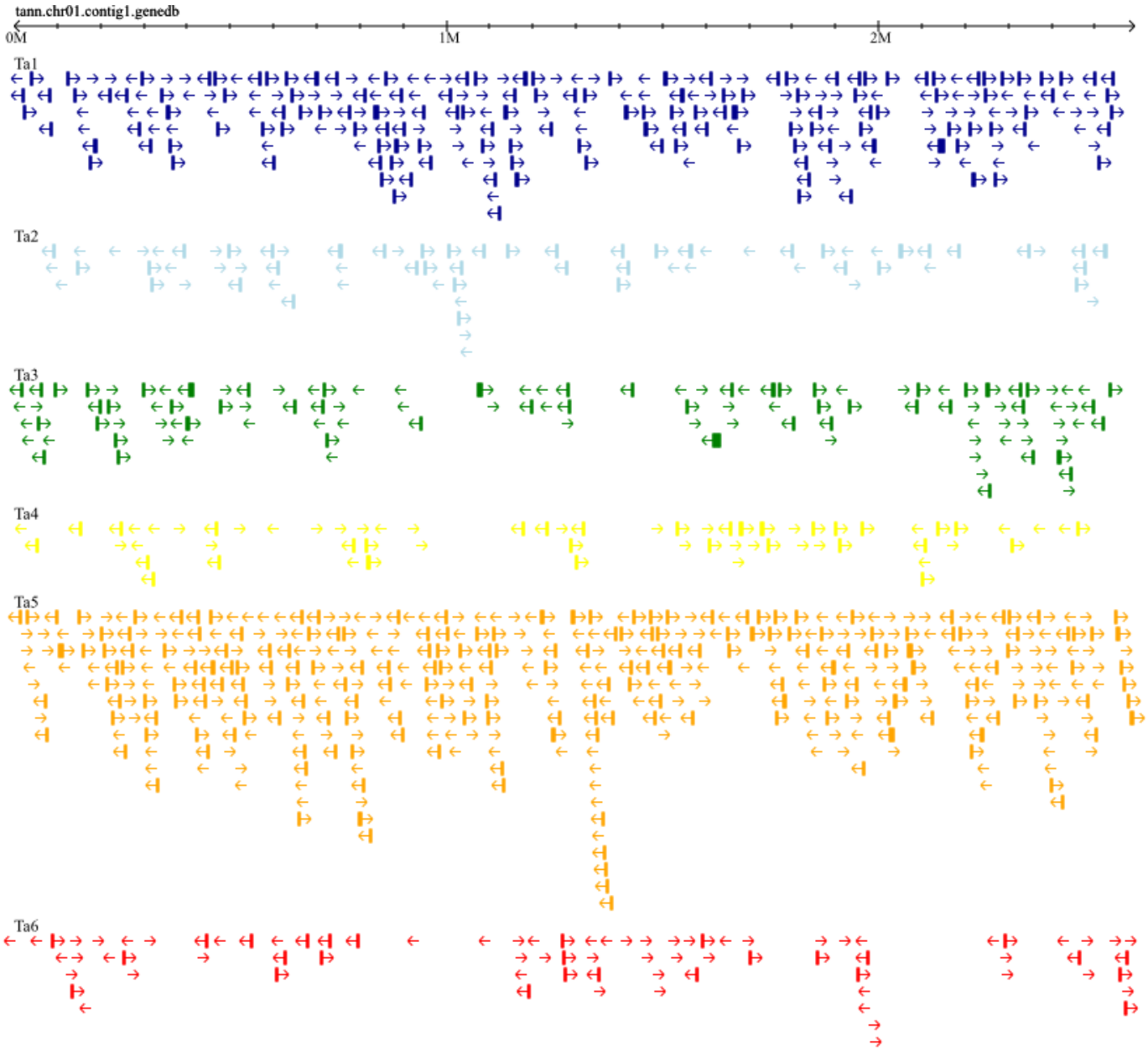

MAL12
0M
1M
2M
Pf1
Pf2
Pf3
Pf4
Pf5
Pf6

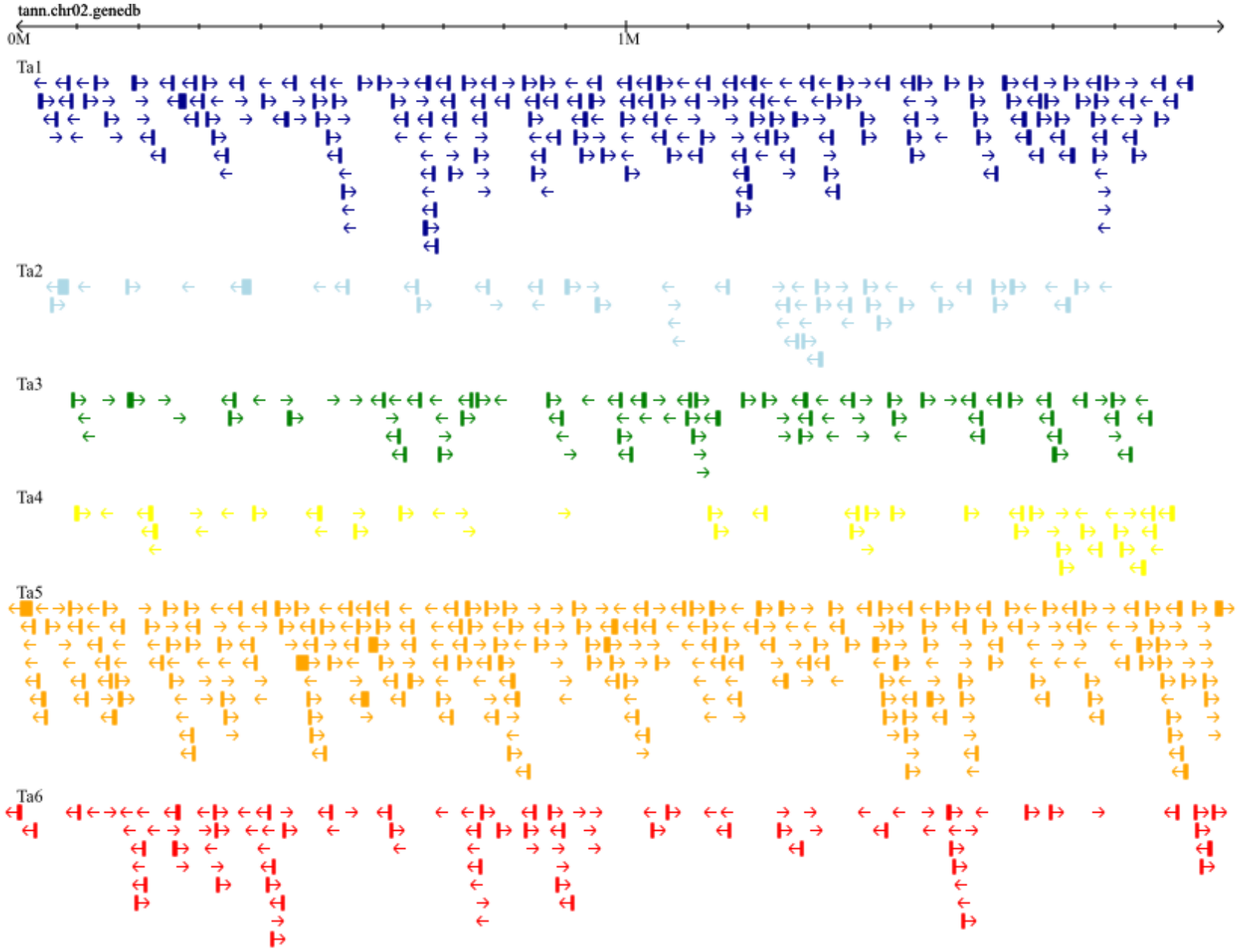

MAL13
0M
1M
2M
Pf1
Pf2
Pf3
Pf4
Pf5
Pf6

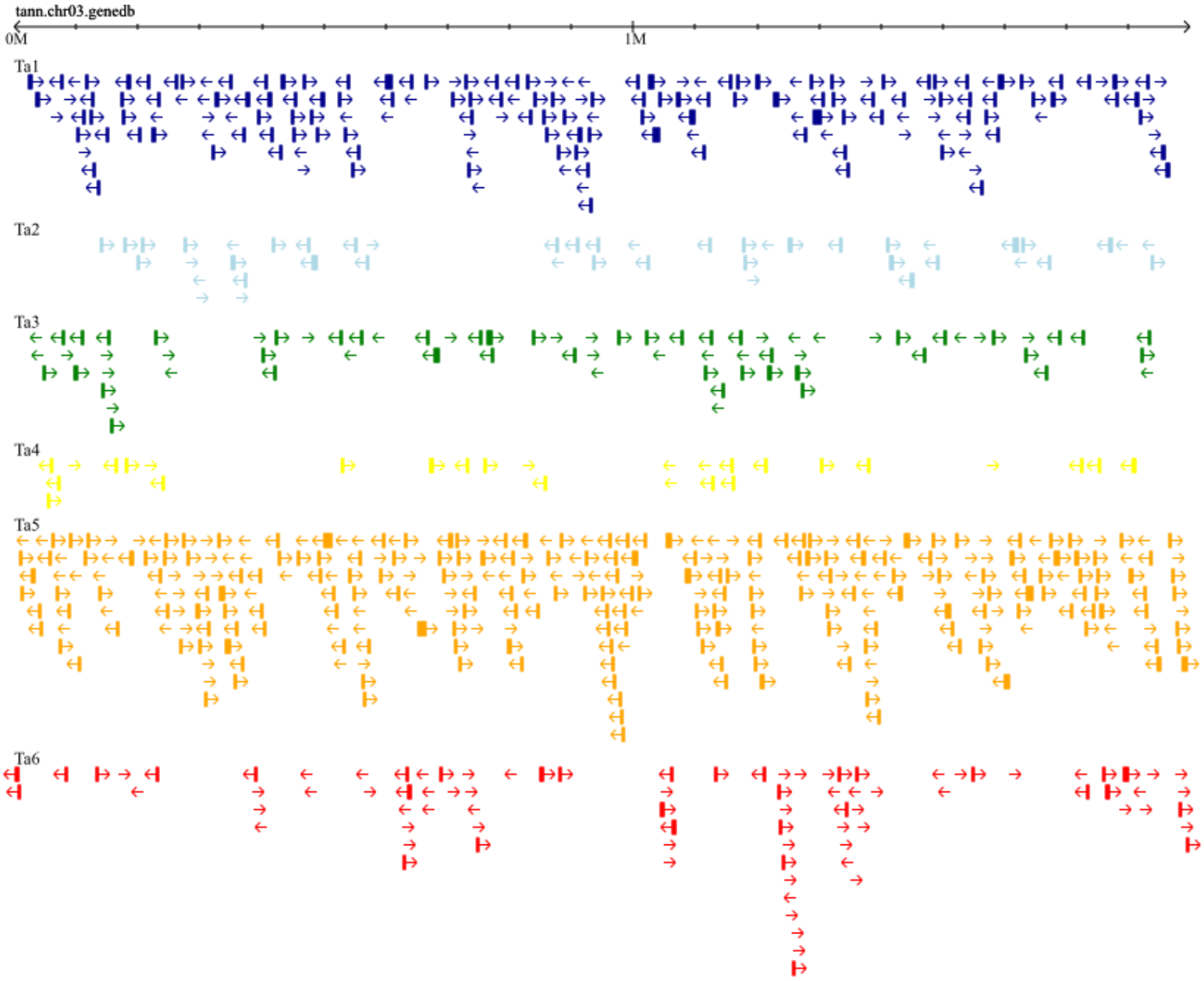

MAL14
0M
1M
2M
3M
Pf1
Pf2
Pf3
Pf4
Pf5
Pf6

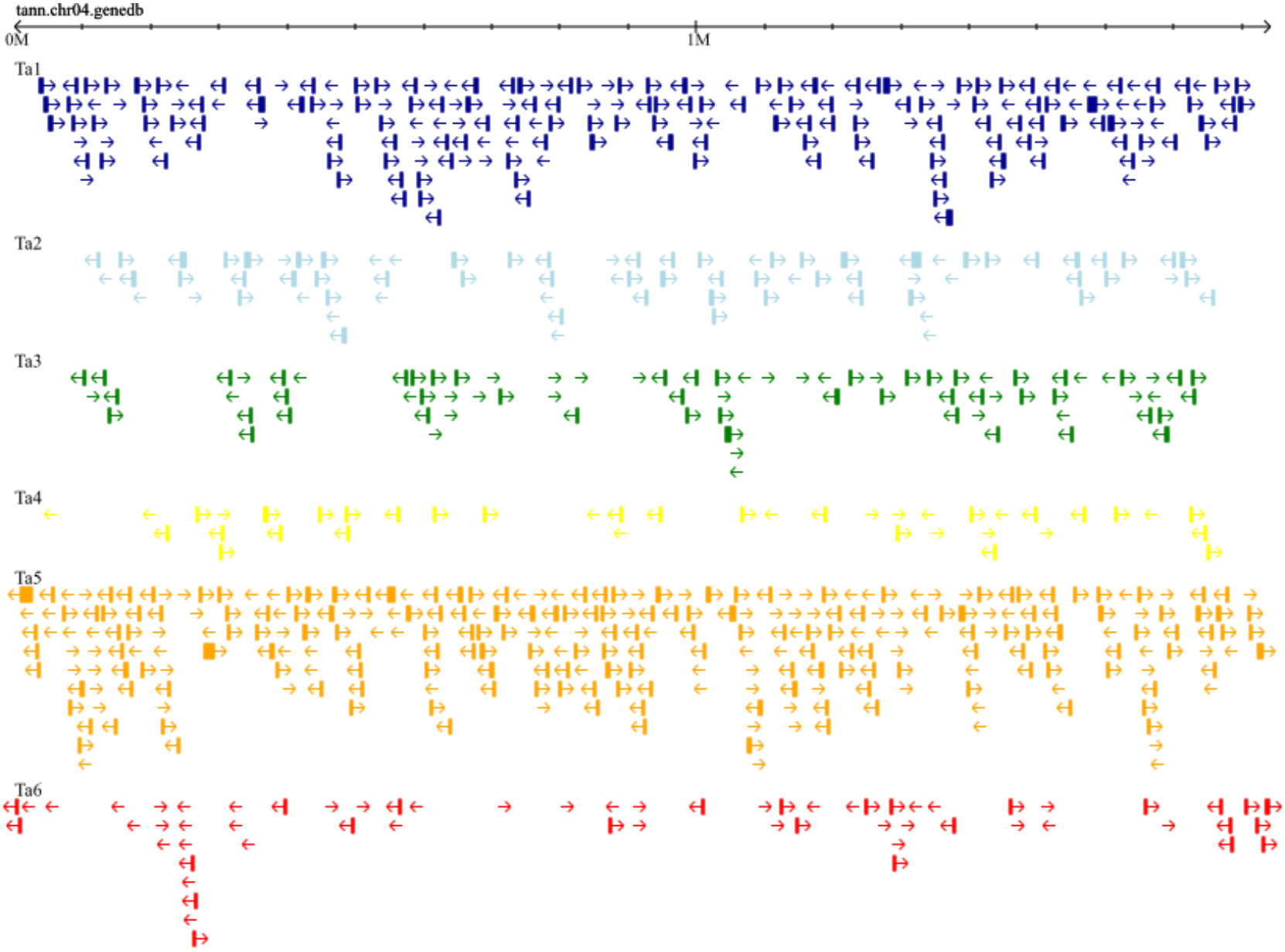
APPENDIX F

CHROMOSOMAL LOCATION OF LINEAGE-SPECIFIC GENES IN *THEILERIA ANNULATA*


tann.chr01.contig1.genedb
0M
1M
2M
Ta1
Ta2
Ta3
Ta4
Ta5
Ta6


tann.chr02.genedb
0M
1M
Ta1
Ta2
Ta3
Ta4
Ta5
Ta6


tann.chr03.genedb
0M
1M
Ta1
Ta2
Ta3
Ta4
Ta5
Ta6


tann.chr04.genedb
0M
1M
Ta1
Ta2
Ta3
Ta4
Ta5
Ta6