5-8-2020

INTEGRATING PHYLOGENOMICS AND CHROMOSOME MAPPING TO STUDY THE EVOLUTIONARY RELATIONSHIPS AMONG EUKARYOTES AND THE EVOLUTION OF THEIR GENOMES

Mario A. Ceron Romero
University of Massachusetts Amherst

INTEGRATING PHYLOGENOMICS AND CHROMOSOME MAPPING TO STUDY THE EVOLUTIONARY RELATIONSHIPS AMONG EUKARYOTES AND THE EVOLUTION OF THEIR GENOMES

A Dissertation Presented

by

MARIO A. CERÓN ROMERO

Submitted to the Graduate School of the
University of Massachusetts Amherst in partial fulfillment
of the requirements for the degree of

DOCTOR OF PHILOSOPHY

May 2020

Organismic and Evolutionary Biology

INTEGRATING PHYLOGENOMICS AND CHROMOSOME MAPPING TO STUDY THE EVOLUTIONARY RELATIONSHIPS AMONG EUKARYOTES AND THE EVOLUTION OF THEIR GENOMES

A Dissertation Presented

by

MARIO A. CERÓN ROMERO

Approved as to style and content by:

_____
Laura A. Katz, Chair

_____
Miguel Mendonça da Fonseca, Member

_____
Li-Jun Ma, Member

_____
Michael Hood, Member

_____
Paige Warren, Program Director
Organismic and Evolutionary Biology

DEDICATION

To Francisca Romero and Eugenio Romero. You were my biggest example of perseverance and constant source of strength, wisdom and love. Every moment spend with you was very inspiring. Not a day goes by that you are not in my heart and my mind.

ACKNOWLEDGMENTS

Thank you Laura for your support and confidence in my work. Your guidance throughout this process has been invaluable. Thank you for letting me be part of the Katz Lab, for teaching me to work in teams, write efficiently, care about the other lab members and love evolution. You have been an incredible mentor.

Thank you Miguel for your support. I learned so much in our collaborations. Working with you in Portugal for a couple of weeks was one of the best experiences I had during my PhD. I hope we keep collaborating and producing amazing papers.

Thank you Li Jun and Michael for your support, encouragement and insightful comments about my research. Also, thank you for your advice about my career plans.

Thank you Katz Lab. You guys were like a family to me in these 5 years and an incredible resource to learn and teach. Thank you very much for all those great memories.

Thank you OEB for your support and for being an amazing program

Thank you to all my friends in the area, especially Diego, Felipe, Matias, Carlos e Itza. You were an important source of emotional and academic support

Thank you my wonderful girlfriend Tatiana Marroquín for your tolerance and support during all this process.

Last, but not least, I want to thank all my family, specially my parents, Raúl Cerón and Nelcy Romero, my sister Evelcy Marcela Cerón, my brother in law César Neira, my aunt

Evelcy Romero. They have always been supporting me and celebrating my achievements. This process would have not been possible without their support.

ABSTRACT

INTEGRATING PHYLOGENOMICS AND CHROMOSOME MAPPING TO STUDY THE EVOLUTIONARY RELATIONSHIPS AMONG EUKARYOTES AND THE EVOLUTION OF THEIR GENOMES

MAY 2020

B.S., UNIVERSIDAD DEL VALLE, COLOMBIA.

M.Sc., UNIVERSIDAD DE PUERTO RICO – RIO PIEDRAS

Ph.D., UNIVERSITY OF MASSACHUSETTS AMHERST

Directed by: Professor Laura A. Katz

Our knowledge about the evolution of eukaryotes and their genomes is very limited because it has largely been based on studies of plants, animals and fungi, which are not a significant representation of the diversity across the eukaryotic tree of life. Advances in sequencing technologies are helping to expand our knowledge by including underrepresented clades and revealing that eukaryotic genomes are much more complex and dynamic than originally thought. In response to the need to explore such levels of complexity in eukaryotic genomes and the earliest events of eukaryotic evolution, this dissertation focuses on the development of bioinformatic and phylogenomic tools to study karyotype evolution and answering deep evolutionary questions. The first chapter covers the development of a phylogenomic chromosome mapper, PhyloChromoMap, and its use to study karyotype evolution in the malaria parasite *Plasmodium falciparum*. In addition to providing a very flexible and powerful tool to map the phylogenetic history of genes across karyotypes, this chapter reveals very distinctive patterns of evolution

between subtelomeric and internal regions of the chromosomes of *P. falciparum*. The second chapter focuses on the development of PhyloToL, a taxon- and gene-rich phylogenomic pipeline. This chapter presents examples of how to use PhyloToL for phylogenomic studies and studies of gene family evolution, and presents a series of benchmark studies comparing PhyloToL against other popular phylogenomic pipelines. Finally, the third chapter focuses on using PhyloToL to explore one of the most critical questions in field of evolution, the root of the eukaryotic tree of life. The results in this chapter suggest that the root should be placed between Opisthokonta and all other eukaryotes. Overall this dissertation contributes insights of the earliest events of evolution in eukaryotes and provides novel approaches to study this topic. The results of this dissertation are important for comparative biology as it allows to understand the timing and mode of evolution of eukaryotic features across the eukaryotic tree of life.

TABLE OF CONTENTS

Page

ACKNOWLEDGMENTS v

ABSTRACT vii

LIST OF TABLES xii

LIST OF FIGURES xiii

CHAPTER

INTRODUCTION 1

1. PHYLOCHROMOMAP, A TOOL FOR MAPPING PHYLOGENOMIC HISTORY ALONG CHROMOSOMES, REVEALS THE DYNAMIC NATURE OF KARYOTYPE EVOLUTION IN *PLASMODIUM FALCIPARUM* 4

    1.1 Abstract 4
    1.2 Introduction 5
    1.3 Material and Methods 8

        1.3.1 Development of PhyloChromoMap 8
        1.3.2 Definition of subtelomeres and detection of young portions and centromeres 9
        1.3.3 Analysis of gene family members: synteny, gene content and dN/dS ratios 9
        1.3.4 Analysis of putative origin of genes 10

    1.4 Results 12

        1.4.1 Development of PhyloChromoMap 12
        1.4.2 Synteny and gene content analyses in young portions 14
        1.4.3 Analysis of SAR-specific and older paralogs 14
        1.4.4 Putative Gene Origin 15

    1.5 Discussion 16

        1.5.1 Patterns of gene conservation in *P. falciparum* and other eukaryotes 16
        1.5.2 Chromosome swapping of subtelomeric regions and evolution of gene families 18
        1.5.3 Putative origin of each gene of *P. falciparum* 19

2. PHYLOTOL: A TAXON/GENE RICH PHYLOGENOMIC PIPELINE TO EXPLORE GENOME EVOLUTION OF DIVERSE EUKARYOTES....26

2.1 Abstract....26
2.2 Introduction....27
2.3 New approaches....31
2.3 Results and discussion....33

2.3.1 Pipeline structure....33
2.3.2 Performance of PhyloToL in GF estimation per taxon....34
2.3.3 Performance of PhyloToL in tree-based contamination removal....35
2.3.4 Implementation for phylogenomic chromosome mapping....37
2.3.5 Test of homology assessment....38

2.4 Materials and methods....40

2.4.1 Naming sequences....40
2.4.2 GF assessment per taxon....41
2.4.3 Refinement of homologs and gene tree reconstruction....43
2.4.4 Tree-based contamination removal....45

3. PHYLOGENOMIC ANALYSES OF 2,700 GENES IN 150 LINEAGES SUPPORT A ROOT OF THE EUKARYOTIC TREE OF LIFE BETWEEN OPISTHOKONTS (ANIMALS, FUNGI AND THEIR MICROBIAL RELATIVES) AND ALL OTHER LINEAGES....54

3.1 Abstract....54
3.2 Introduction....55
3.3 Results....58

3.3.1 Building the phylogenomic datasets....58
3.3.2 Inference on location of the root....59
3.3.3 Comparison to published hypotheses....60

3.4 Discussion....61
3.5 Methods....65

3.5.1 Taxa selection....65
3.5.2 Gene family selection....65
3.5.3 Root inference....66
3.5.4 Comparing different root hypotheses....67

APPENDICES

A. SUPPLEMENTARY TABLES....71
B. SUPPLEMENTARY FIGURES....78

BIBLIOGRAPHY.....90

LIST OF TABLES

Table Page

1.1. Summary of conservation of genes in *P. falciparum*.....20

2.1. Summary of the experiment of gene family assessment per taxon.....46

2.2. Summary of conservation of genes in *Trypanosoma brucei* 47

3.1. A summary of taxon selection for each dataset 68

S1. Size of young region in chromosomes of *P. falciparum* 72

S2. Characteristics of putative centromeres in chromosomes of *P. falciparum* 73

S3. Genes of *P. falciparum* that were likely transferred through interdomain LGT 74

S4. Comparison of features among PhyloToL, OneTwoTree (Drori, et al. 2018), SUPERSMART (Antonelli, et al. 2017) and PhyloTA (Sanderson, et al. 2008). 75

S5. PhyloToL homology test for candidate superfamilies proposed by Reddy and Saier (2016).....76

S6. Statistical comparison of Fungi + others root against previously published roots using t-student test.....77

LIST OF FIGURES

Figure Page

1.1. Exemplar phylogenomic maps of chromosomes 1, 2 and 7 of *Plasmodium falciparum* 3D7 highlighting 'young' subtelomeric and internal regions (boxes) 21

Figure 1.2. Exemplar phylogenomic maps of chromosomes 1-3 of *Saccharomyces cerevisiae* S288C 22

Figure 1.3. Paralogs in a) subtelomeric regions of *P. falciparum* 3D7 tend to be young while paralogs in b) internal regions tend to be old.....23

Figure 1.4. Paralogs from gene family *var* (blue) do not exhibit significant differences in selection intensity (i.e. dN/dS) according to location, while paralogs from other gene families (red and black) show significant differences between subtelomeric and internal regions.....24

1.5. Exemplar phylogenomic map of the chromosomes 1, 2 and 7 according to the hypothetical origin of genes.....25

2.1. The four components of PhyloToL.....48

2.2. Evaluation of performance of the first and second component of PhyloToL.....49

2.3. Examples of contamination from gene trees, which are used to define rules for the contamination removal loop of component 3 of PhyloToL.....50

2.4. Example of phylogenomic map of the chromosome III of *Trypanosoma brucei* generated by combining PhyloToL and PhyloChromoMap 51

2.5. PhyloToL homology assessment for well-known GFs that duplicated prior to LUCA. Subfamilies of these ancient GFs are often categorized in different orthologous groups by OrthoMCL.....52

2.6. PhyloToL homology assessment for candidate superfamilies (S) of outer membrane pore-forming proteins as proposed by Reddy and Saier (2016) 53

3.1. A root between fungi and all other eukaryotes is the most parsimonious hypothesis inferred from 100 iterations for each of our four datasets 69

3.2. Comparison of five hypotheses for the root from the literature estimated using iGTP with the 4 datasets (repetitions) 70

S1. Flow diagram of the methods for mapping the chromosomes of *P. falciparum* with PhyloChromoMap.....79

S2. Phylogenomic map of chromosomes of *Plasmodium falciparum* 3D7 showing the conservation level of genes assessed 80

S3. Phylogenomic map of chromosomes of *Saccharomyces cerevisiae* S288C showing the conservation level of genes assessed 81

S4. Analysis of synteny shows that synteny blocks are not shared between internal young regions (white boxes) and subtelomeric young regions.....82

S5. Genes in young regions tend to be restricted to either subtelomeric or internal regions, with the exception of *var* genes that are abundant in both subtelomeric and young regions 83

S6. Subtelomeric and internal paralogs of *var* genes do not have significant differences in their dN/dS ratios 84

S7. Phylogenomic map of the chromosomes of *P. falciparum* according to the hypothetical origin of genes.....85

S8. Detailed phylogenomic map of the chromosomes of *Trypanosoma brucei* generated by combining outputs of PhyloToL with PhyloChromoMap 86

S9. High levels of conservation of many genes across chromosomes (thick lines) of *Trypanosoma brucei* 87

S10. Genome size comparison between the Metazoa and Fungi.....88

S11. Number of trees with at least three species per minor clade in dataset SEL+.....89

INTRODUCTION

Historically, our knowledge about the evolution of eukaryotes and their genomes has largely been based on studies of plants, animals and fungi, limiting our view of the earliest evolutionary events. These analyses led to the view of the eukaryotic genomes are static entities with fixed karyotypes. Things have been changing dramatically with the advances on molecular tools including high throughput sequencing platforms (e.g. 454, Illumina, PacBio) that allow more efficient exploration of genomes. Many new clades have been described, which is critical as the bulk of the diversity on eukaryotes lies out of animals, plants and fungi. These studies have found that eukaryotic genomes are more dynamic than the canonical view (McGrath and Katz 2004; Zufall, et al. 2005; Parfrey, et al. 2008). Then, advances in sequencing technologies offer the possibility to explore deep evolutionary concerns in eukaryotic history, such as the factors that drive karyotype evolution and the reconstruction of the oldest phylogenic relationships. High throughput sequencing technologies also come with challenges. For instance, sequence contamination, bioinformatic errors in annotations and evolutionary events (e.g. lateral gene transfer, incomplete lineage sorting) affect phylogenetic inferences.

Given this background, the three chapters of this dissertation have two broad aims. First, the development of bioinformatic tools for phylogenomic and chromosome mapping analyses that account on the challenges of the high throughput sequencing technologies. Second, the implementation of those bioinformatic tools to study deep issues in eukaryotic evolution: the evolution of karyotypes and the root of the eukaryotic tree of life. Together, these three chapter will challenge our view the earliest events in the

evolution of eukaryotes as well as contribute to the study of karyotype evolution in eukaryotes.

The first chapter focuses on the development of PhyloChromoMap (Ceron-Romero, et al. 2018), a tool for mapping the evolutionary history of genes across the chromosomes. PhyloChromoMap requires a physical map of the chromosomes and a set of gene trees. The main goal of PhyloChromoMap is to estimate the level of conservation in gene trees based on presence/absence of taxa, and display it in the physical map. This chapter also presents the phylogenomic map of the chromosomes of *Plasmodium falciparum*, the causative agent of malaria in Africa, as an exemplary case to show the uses of PhyloChromoMap. Although previous research predicted that subtelomeric regions are highly dynamic in *P. falciparum* (Freitas-Junior, et al. 2000b; Scherf, et al. 2001; Hernandez-Rivas, et al. 2010), this is the first time that this is demonstrated integrating genomic and phylogenomic data with chromosome mapping information.

The second chapter of this dissertation is dedicated to the improvement of PhyloToL (Ceron-Romero, et al. 2019), a custom phylogenomic pipeline. PhyloToL is the last version of the previously published Katzlab phylogenomic pipeline (Grant and Katz 2014a). This chapter focuses on the improvements that were made for creating this last version and discussing the features that make PhyloToL to stand up among other phylogenomic pipelines. Some of these features are: flexibility/modularity, capability to integrate data from different sources (i.e. genomes, transcriptomes and protein data), efficiency to detect and remove sequence contamination and support of a wide range of diversity (including ~2 million years old relationships). Along with the improvement of PhyloToL, technical evaluation, and benchmark studies, this chapter contains an analysis

of the phylogenomic map of the chromosomes of *Trypanosome brucei*, the sleeping sickness parasite, as an example of integration of PhyloToL and PhyloChromoMap (Chapter 1).

The third chapter focuses on estimating the most likely root of the eukaryotic tree of life (i.e. EToL) using PhyloToL (Chapter 2) in combination with a gene tree – species tree reconciliation method. This approach estimates the species tree that requires the fewest duplications and losses to explain the topology of a set of gene trees (gene tree parsimony, Guigo, et al. 1996). The key difference from the supermatrix method, the most common method in studies about the root of EToL, is that it takes advantage of the phylogenetic signal of paralogs instead of removing them for further concatenation. The result of this analysis predicts that the root should be placed either between Opisthokonta (i.e. animals and fungi) and the others or between Fungi and the other. The discussion also includes a section explaining how a root between Fungi and the others could be an artifact caused by high rates of gene loss in Fungi. The results of this research contradict the current 'popular' views of either a unikont-bikont or Excavata root.

Overall this dissertation furthers our understanding of the immense diversity on earth and the complexity of the eukaryotic genomes. More specifically, this dissertation provides insights of the earliest events of evolution of the eukaryotic genomes and provides novel approaches to study this topic. The work here will allow to have a better sense of what characters are ancestral in eukaryotes. Also, this work will promote the study of deep phylogenetic questions in eukaryotes and the study of karyotype evolution in other eukaryotic lineages.

CHAPTER 1

PHYLOCHROMOMAP, A TOOL FOR MAPPING PHYLOGENOMIC HISTORY ALONG CHROMOSOMES, REVEALS THE DYNAMIC NATURE OF KARYOTYPE EVOLUTION IN *PLASMODIUM FALCIPARUM*[1]

1.1 Abstract

The genome of *P. falciparum*, the causative agent of malaria in Africa, has been extensively studied since it was first fully sequenced in 2002. However, many open questions remain, including understanding the chromosomal context of molecular evolutionary changes (e.g. relationship between chromosome map and phylogenetic conservation, patterns of gene duplication, and patterns of selection). Here we present PhyloChromoMap, a method that generates a phylogenomic map of chromosomes from a custom-built bioinformatics pipeline. Using *P. falciparum* 3D7 as a model, we analyze 2116 genes with homologs in up to 941 diverse eukaryotic, bacterial and archaeal lineages. We estimate the level of conservation along chromosomes based on conservation across clades, and identify 'young' regions (i.e. those with recent or fast evolving genes) that are enriched in subtelomeric regions as compared to internal regions. We also demonstrate that patterns of molecular evolution for paralogous genes differ significantly depending on their location as younger paralogs tend to be found in subtelomeric regions while older paralogs are enriched in internal regions. Combining these observations with analyses of synteny, we demonstrate that subtelomeric regions

[1] Cerón-Romero MA, Nwaka E, Owoade Z, Katz LA. 2018. PhyloChromoMap, a tool for mapping phylogenomic history along chromosomes, reveals the dynamic nature of karyotype evolution in *Plasmodium falciparum*. *Genome Biol Evol.* 10:553-561.

are actively shuffled among chromosome ends, which is consistent with the hypothesis that these regions are prone to ectopic recombination. We also assess patterns of selection by comparing dN/dS ratios of gene family members in subtelomeric vs internal regions, and we include the important antigenic gene family *var*. These analyses illustrate the highly dynamic nature of the karyotype of *P. falciparum*, and provide a method for exploring genome dynamics in other lineages.

1.2 Introduction

Numerous studies of plants, animals and fungi have formed our classical view of karyotypes as stable entities that have only minor variations within species (Hope 1993; Sites and Reed 1994; Schubert and Vu 2016). However, an increasing number of studies of unicellular eukaryotes in last decades has revealed that karyotypes are more dynamic than originally thought (McGrath and Katz 2004; Zufall, et al. 2005; Parfrey, et al. 2008; Katz 2012; Oliverio and Katz 2014). For instance, recombination between non-homologous chromosomes (i.e. ectopic recombination) can lead to intraspecific variation of the karyotype in the model organism *Saccharomyces cerevisiae* (Loidl and Nairz 1997). In parasites such as *Giardia lamblia*, *Encephalitozoon cuniculi* (Biderre, et al. 1999) and *Encephalitozoon hellem* (Delarbre, et al. 2001) and *Plasmodium falciparum* (Freitas-Junior, et al. 2000a; Scherf, et al. 2008; Hernandez-Rivas, et al. 2013; Claessens, et al. 2014) the same type of chromosomal rearrangements contributes to antigenic variation, which allows escape from the host immune system. Most of these karyotype variations have been described using microscopy and/or analyses of limited sets of genes (Loidl and Nairz 1997; Biderre, et al. 1999; Freitas-Junior, et al. 2000a; Delarbre, et al. 2001).

The growing number of genomes that are available enables the development of methods to explore patterns of karyotype evolution. Well-annotated genomes can be used to build physical maps in order to compare structural characteristics such gene content and synteny. For instance, genome maps have allowed detection of differences in synteny among species of the lineages *Ostreococcus* (Palenik, et al. 2007), *Plasmodium* (Carlton, et al. 1999; Kooij, et al. 2005), *Saccharomyces* (Walther, et al. 2014), *Trypanosoma* (Ghedin, et al. 2004). Likewise, for phylogenomic analyses, the increase in genomic data provides more taxa and genes to compare. Yet, analysis of the phylogenetic history of genes along chromosomes can yield important insights about the evolution of karyotypes.

Plasmodium falciparum, the most virulent of the human malaria parasites, is a good model to study karyotype evolution because its life cycle has been extensively studied and its genome has been fully sequenced (Gardner, et al. 1998; Gardner, et al. 2002). The AT-rich genome of *P. falciparum* is divided among 14 chromosomes that harbor housekeeping genes in their internal regions and antigen genes at their ends (Gardner, et al. 2002). Because of the importance of antigenic variation as *P. falciparum* evades host immune system, the ends of the chromosomes (which are enriched for antigenic gene families) have been relatively well characterized (de Bruin, et al. 1994; Pace, et al. 1995). In *P. falciparum*, these regions are marked by telomeres, followed by a ~40 kb region, the 'telomere associated sequences', that contains a series of repeat sequences (Figueiredo, et al. 2000; Figueiredo, et al. 2002; Figueiredo and Scherf 2005; Hernandez-Rivas, et al. 2013). Antigen genes *var*, *rif* and *stevor* are located after 40 kb, where the abundance of repeated genes makes this region prone to ectopic recombination (Scherf, et al. 2001; Hernandez-Rivas, et al. 2013). This observation has led to the

proposal that subtelomeric regions in *P. falciparum* evolve through ectopic recombination between chromosomes (Freitas-Junior, et al. 2000a; Scherf, et al. 2001; Hernandez-Rivas, et al. 2013).

Genomes from other apicomplexans have been completed, enabling comparative genomic analyses between those lineages and *P. falciparum.* Previous studies comparing presence and absence of genes show high conservation in gene content among *Plasmodium* species (Carlton, et al. 2002; Carlton, et al. 2008; Pain, et al. 2008). While comparisons among apicomplexan species revealed that few genes are shared among all species (<34%; Kuo, et al. 2008; Kissinger and DeBarry 2011).

We decided to explore further the evolution of the *P. falciparum* genome by analyzing the phylogenetic conservation of genes and gene families in their chromosomal context. In order to achieve this goal, we develop a method, PhyloChromoMap*,* to depict the evolutionary history of genes along a chromosomal map. Using *P. falciparum* as a case of study we infer the phylogeny of its genes with a taxon-rich phylogenomic pipeline (Grant and Katz 2014a; Katz and Grant 2015). Then, we estimate the level of conservation of protein coding sequences by determining the presence or absence of homologs in other clades (i.e. Bacteria, Archaea, Opisthokonta, Archaeplastida, SAR, Excavata, Amoebozoa and other eukaryote lineages) in single gene trees. We also assess patterns of molecular evolution in paralogs across chromosomes, and provide a map that indicates putative origin of genes.

1.3 Material and Methods

1.3.1 Development of PhyloChromoMap

Starting from a phylogenomic pipeline previously developed in our lab (Grant and Katz 2014a; Katz and Grant 2015), we created PhyloChromoMap to map the evolutionary history of genes along chromosomes (https://github.com/Katzlab/PhyloChromoMap_py). Our initial collection of homologs uses gene families defined in OrthoMCL (http://www.orthomcl.org/orthomcl/) and as such, each of these clusters of homologs is referred to as an "orthologous group" or OG. We analyze a total of 5336 putative coding genes from *P. falciparum* 3D7 (assembly ASM276v1) by BLAST (Altschul, et al. 1990) against OrthoMCL (Figure S1). This results in 2116 genes falling in 1962 OGs that are represented in our pipeline. The remaining OGs are not represented in our pipeline either because they contain very few homologs or because they produce very poor-quality alignments that are discarded in subsequent steps of the pipeline; these are labeled as NIP (not in pipeline) in tables and figures. We represent graphically the number of minor clades (e.g. Apicomplexa) per major clade (e.g, SAR) for every OG in our pipeline (Figures 1.1, S1, S2). We then use the R "image" function (Team 2016), which uses a matrix to display spatial data, to display the phylogenomic history of genes along the chromosome map. In order to validate our method and results for *P. falciparum*, we implemented PhyloChromoMap also in the model organism *Saccharomyces cerevisiae* S288C (Figures 1.2, S3).

1.3.2 Definition of subtelomeres and detection of young portions and centromeres

We defined subtelomeric regions after producing the chromosome maps and observing that all chromosome ends contain well defined young regions. We then focus on subtelomeric regions that contain the most distal 15% of the chromosome or the final 200 kb (whichever is smaller) to capture these young regions. We use a custom Ruby script to walk the chromosomes and detect young portions in the subtelomeric and internal regions (Figure S1). Young portions are regions in which genes are in less than 3 major eukaryotic clades, though we allow the presence of one gene conserved in 3 or more major clades. Moreover, we illustrate a gene as present in a major clade only if it is found in at least 25% of its minor clades to account for spurious results and intradomain Lateral Gene Transfer (LGT). We searched young portions in both subtelomeric and internal, only considering internal young regions that are ≥ 90 kb (Table S1). All chromosomes except chromosome 10 have a region of around 2-3 kb with the highest GC content, 94-98%. This region is assumed as centromere (Bowman, et al. 1999; Hall, et al. 2002). In chromosome 10 this region is less obvious, encompassing only around 1 kb with a 94% GC content (Table S2).

1.3.3 Analysis of gene family members: synteny, gene content and dN/dS ratios

We perform a synteny analysis of subtelomeric and internal young portions using SyMAP (Soderlund, et al. 2006) (Figure S1). We explore different values for the minimum number of anchors to define a synteny block (i.e. from 3 to 7) and do not see any major differences (Figure S4). We choose parameters to better retain duplications: N=2 (retain the anchors with scores among the top 2) and anchor scores ≥ 80% of the second best anchor. Finally, overlapping synteny blocks are merged. We also survey the

gene content of young portions, including *Plasmodium* specific coding domains (Figure S1). We categorize the sequences by gene family when possible and plot their frequency as a heatmap (Figure S5).

We use CIRCOS plots (Krzywinski, et al. 2009) to map paralogs of genes that match OGs (Figures 1.3, S1). In CIRCOS, we choose the links option for representing these paralogs, with a single link connecting each pair of paralogs. The relative age of paralogs is calculated as the number of major clades that contain them and is also displayed in the plots. Additionally, pairwaise dN/dS values are calculated for all paralogs using yn00, PAML (Yang 1997) and compared between subtelomeric and internal paralogs (Figure 1.4).

We conduct a phylogenetic analysis for protein sequences of *var* using RAxML (Stamatakis 2014) and model of evolution WAG+I+G+F. The model of evolution is inferred using Prottest3 (Darriba, et al. 2011). The resulting phylogenetic tree is used to calculate a dN/dS value (free ratio model) using codeML-PAML (Yang 1997) and HyPhy (Kosakovsky Pond, et al. 2005) (Figure S6). Difference of selection intensity between internal and subtelomeric copies is analyzed using the software RELAX from the Datamonkey package (Wertheim, et al. 2015). This analysis is not performed in other antigenic gene families such as *rif* and *stevor*, because there are few *rif* and no *stevor* paralogs in the internal regions of the chromosomes.

1.3.4 Analysis of putative origin of genes

We use two approaches to detect both recent and old interdomain LGT event in *P. falciparum*, a parametric approach based on nucleotide composition and a phylogenetic

approach (Table S3). For the parametric approach, we calculate the average GC content per chromosome and per gene; when the average GC content in a gene is two standard deviations away from the chromosomal average GC content, the gene is considered as a candidate laterally transferred gene. Then, we use BLAST to assess whether the gene is shared only between Apicomplexa and prokaryotes. For the phylogenetic approach, we explore the topology of gene trees with custom python scripts that incorporate P4, a maximum likelihood and Bayesian package (Foster 2004). In the topology of the gene trees, we identify potential interdomain LGTs when: (i) the gene trees contain only prokaryotes and Apicomplexa; and (ii) Apicomplexa lineages are monophyletic and nested or sister to a clade of Bacteria/Archaea.

We also estimate putative origin of genes by counting presence and absence of taxa in gene trees. Archaea, Bacteria or major clades of Eukaryotes are considered as present in a gene tree if at least 25% of their minor clades are present. Genes that have bacteria and at least 5 of the eukaryotic major clades (considering orphans ("EE" – everything else) as a major clade) are candidate Endosymbiotic Gene Transfers (EGTs) from mitochondria. Genes that have bacteria at least 2 major clades of photosynthetic eukaryotes (i.e. SAR, Archaeplastida, some orphans) are candidate EGTs from the plastid. Genes that have at least 5 eukaryotic major clades and no prokaryotes are candidate conserved genes from the Last Eukaryotic Common Ancestor (LECA). Genes present in Archaea and at least 5 eukaryotic major clades are candidate conserved genes from the Last Archaeal Common Ancestor (LACA, which includes the ancestor of eukaryotes (Williams, et al. 2013; Hug, et al. 2016)). Finally, genes present in Archaea,

Bacteria and at least 5 eukaryotic major clades have a putative origin in the Last Universal Common Ancestor (LUCA). All these genes were mapped (Figures 1.5, S7).

1.4 Results

1.4.1 Development of PhyloChromoMap

We built PhyloChromoMap to map the evolutionary history of genes along chromosomes, and we use *Plasmodium falciparum* as a test case. In sum, we started with a collection of 13104 multisequence alignments generate in Guidance (Sela, et al. 2015a) and corresponding gene trees built in RaxML (Stamatakis 2014), which included up to 519 Eukaryotes, 303 Bacteria and 119 Archaea (Grant and Katz 2014a; Katz and Grant 2015). PhyloChromoMap estimates the phylogenetic conservation for every gene based on the presence/absence of major and minor lineages in single gene trees (See methods, Table 1.1). We then use function "image" in R (Team 2016) to map the phylogenetic conservation of each gene along each chromosomes.

We use PhyloChromoMap to estimate the level of conservation of 5,336 protein coding genes along the chromosomes of *P. falciparum* strain 3D7. The results indicate that 21% of the genes of *P. falciparum* are present in at least some representatives of all major eukaryotic clades (i.e. SAR, Archaeplastida, Excavata, Amoebozoa, and Opisthokonta; Table 1.1). Some genes are more ancient/conserved as they are also shared with Archaea (3%), Bacteria (4%) or both Archaea and Bacteria (5%). In contrast, 2% of the genes are more recent as they are present only in *Plasmodium* and other members of the SAR clade. Roughly 60% of 'genes' (i.e. ORFs) in the *P. falciparum* genome are fast evolving, unique to *Plasmodium* and/or are mis-annotated; this group of genes are

considered ‘not in pipeline’ (NIP) in our analyses as they do not pass our criteria for generation of multisequence alignments and trees (see methods).

We built phylogenomic maps of the 14 chromosomes of *P. falciparum* 3D7 to illuminate patterns of conservation across different chromosomal regions (Figures 1.1, S2). Distinct patterns of conservation are found across chromosomes. For instance, while internal regions contain primarily conserved genes (i.e. genes with many homologs in other lineages), subtelomeric regions contain almost exclusively young genes. We recognize that ‘young’ genes will include both fast evolving genes (i.e. those whose identity to homologs is very low) as well as genes with recent origins. We determine the length of ‘young’ regions (i.e. those containing genes shared with members of two or fewer major eukaryotic clades, allowing for a single ‘interrupting’ gene) and found that subtelomeric young regions average 134 kb (range of 85-218 kb; Table S1), and internal young regions average 106 kb (range of 91 -141 kb; Table S1). On the other hand, centromeric regions do not exhibit any clear pattern of gene conservation as these regions harbor young genes in some chromosomes (e.g. chromosomes 3 and 7) and old/conserved in others (e.g. chromosomes 2 and 5; Figures 1.1, S2).

To exemplify further the power of Phylochromomap*,* we also generated the phylogenomic map of the chromosomes of *S. cerevisiae* in order to validate our method (Figures 1.2, S3). Overall this map shows a higher density of genes than we observe for *P. falciparum* and here too we do not see any pattern of near the centromeres (Figures 1.2, S3). Unlike the pattern for *P. falciparum,* we find no evidence of young subtelomeric regions except for chromosome I, which contains a dense central region flanked by low gene density in the distal regions (Figure 1.2). Previous studies reveal that chromosome I

is rich in rRNA genes (Seligy and James 1977) and unexpressed pseudogenes, suggesting that these regions represent the yeast equivalent of heterochromatin (Bussey, et al. 1995).

1.4.2 Synteny and gene content analyses in young portions

We test for recombination between subtelomeric (ST) regions and internal (IN) young portions of chromosomes through analysis of synteny (Figure S4) and comparison of gene content (Figure S5). Chromosomes share blocks of sequences in conserved order (i.e. synteny blocks) in subtelomeric regions (ST) with a few exceptions (14ST3', 14ST5', 5ST3' and 11ST3'; Figure S4). Some subtelomeric regions (e.g. 13ST3', 1ST5', 11ST5') have complex patterns of synteny, with many bocks shared with other subtelomeric regions. In contrast, internal young regions (IN) do not share synteny blocks. In addition, although there are some gene family members shared between young portions of internal and subtelomeric regions, subtelomeric regions tend to harbor more antigenic genes such as *var*, *rif*, and *stevor* (Figure S5).

1.4.3 Analysis of SAR-specific and older paralogs

We compare the patterns of evolution of gene family members across subtelomeric and internal regions of the chromosomes. We analyze both levels of conservation and selection intensity, the latter estimated by dN/dS ratios (Yang 1997; Kosakovsky Pond, et al. 2005; Wertheim, et al. 2015). Maps of subtelomeric and internal paralogs demonstrate that while subtelomeric regions tend to accumulate more 'young' or SAR-specific paralogs, internal regions tend to accumulate 'old' paralogs that are conserved in five or more major clades (Figure 1.3). There is also a difference in the patterns of selection acting on subtelomeric and internal paralogs: subtelomeric paralogs

tend to have higher and more variable dN/dS ratios (mean 0.48, 95% CI 0.42-0.53) than paralogs in internal regions (mean 0.15, 95% CI 0.13-0.16). This implies that paralogs in internal regions are more consistently subject to functional constraint than subtelomeric paralogs.

Paralogs of the gene family *var,* which encode for PfEMP1 antigens, exhibit different patterns than paralogs of other genes. The *var* genes are young as they are specific of *P. falciparum* and are also frequently found in internal regions (Figures 1.1, S4). Moreover, dN/dS ratios are relatively high for *var* genes (mean 0.5, 95% CI 0.46-0.54) (Figures 1.4, S6). In contrast to patterns for other gene families, there are no significant differences among dN/dS ratios between internal and subtelomeric *var* paralogs based on RELAX, a hypothesis testing framework for detecting relaxed selection (Wertheim, et al. 2015). This suggests that natural selection coupled with recombination contributes to levels of variation among *var* genes, which in turn are important in enabling these parasites to escape host immune systems (Kyes, et al. 2007).

1.4.4 Putative Gene Origin

Given that our novel method connects the physical chromosomal map with the evolutionary history of genes sampled from across the tree of life, we can map putative origins of genes along chromosome maps. Using an approach based on differences of GC content, we detect one possible case of a recent interdomain LGT event involving *P. falciparum* and prokaryotes (Table S3). This gene (FIRA) is an interspersed repeat antigen, which is involved in drug resistance (Stahl, et al. 1987). Moreover, analyzing single gene trees, we detect 9 possible cases of ancient LGT events involving prokaryotes

and Apicomplexa (Table S3). Here we identify cases where apicomplexan sequences are nested within bacterial clades in single gene trees (see methods). These genes have varied function and do not display any distinctive pattern of distribution in the chromosomes (Figure S2).

We also assign genes along our chromosome map to categories of putative origins, which can then be used for further investigation. For example, genes that are widely distributed in bacteria, archaea and eukaryotes may date to LUCA while genes found only in photosynthetic eukaryotes (and sometimes also some bacteria) may represent cases of EGT from plastids (Figures 1.5, S7). Based on an analysis of presence/absence of taxa on gene trees, we detected 179 genes that are candidate cases of EGT from plastids and 148 genes that are candidate cases of EGT from mitochondria (or bacteria). We also detected 844 genes that are maybe conserved from LECA, 151 from LACA and 238 putatively from LUCA (Figures 1.5, S7).

1.5 Discussion

1.5.1 Patterns of gene conservation in *P. falciparum* and other eukaryotes

Here we present PhyloChromoMap, a novel method that combines the power of phylogenomics and genome mapping to explore patterns of karyotype, gene and molecular evolution. Using *P. falciparum* as a model, we characterize the level of evolutionary conservation in genes along all fourteen chromosomes. This analysis demonstrates that subtelomeric regions are young as compared to internal chromosome regions, which contain a mixture of conserved and lineage-specific genes (Figures 1.1, S2). These data, and the evidence of syntenic blocks among subtelomeres (Figure S4), are

consistent with the hypothesis that chromosomes of *P. falciparum* are actively swapping subtelomeric regions due to frequent ectopic recombination (Freitas-Junior, et al. 2000a; Scherf, et al. 2001; Scherf, et al. 2008; Hernandez-Rivas, et al. 2013). Analyses using fluorescent *in situ* hybridization reveal that chromosomes of *P. falciparum* attach to the nuclear periphery in clusters, suggesting that these clusters may facilitate recombination across subtelomeric regions of chromosomes (Freitas-Junior, et al. 2000a).

Differences in levels of conservation across chromosomes exist in diverse lineages from across the tree of life. For instance, the soil bacterium *Streptomyces* also has more conserved genes in the internal part of their linear chromosomes and the younger genes towards chromosome ends (Bentley, et al. 2002; Ikeda, et al. 2003; Chater 2016). As is the case for *P. falciparum,* young genes in *Streptomyces* evolve by recombination, mostly with linear plasmids or segments of chromosomes from other *Streptomyces* (Chater 2016). Other eukaryotic lineages such as the yeast *Saccharomyces* and the parasites *Giardia intestinalis* and *Encephalitozoan cuniculi* also tend to have younger genes toward the chromosome ends (Kellis, et al. 2003; Ankarklev, et al. 2015; Dia, et al. 2016). Chromosome ends in these lineages are also subject to rearrangements such translocations or duplications, which promotes diversity in telomeric and subtelomeric gene families (Kellis, et al. 2003; Ankarklev, et al. 2015). In contrast, the highly conserved ribosomal DNA loci are found in subtelomeric regions of the nucleomorph (remnant nuclei from algal symbionts) genomes in cryptomonads and chlorarachniophytes (Lane and Archibald 2006; Lane, et al. 2006; Silver, et al. 2010; Tanifuji, et al. 2014).

1.5.2 Chromosome swapping of subtelomeric regions and evolution of gene families

We analyze the relationship between level of conservation of duplicated genes and chromosomal location, and find that paralogs in subtelomeric regions tend to be young as compared to those throughout the rest of the chromosome map (Figure 1.3). Mechanisms underlying gene duplication in eukaryotes include unequal crossing over, transposition/retrotransposition and genome or segmental duplication (Hahn 2009). The use of PhyloChromoMap reveals that gene duplication occurs during the shuffling of subtelomeric regions between chromosomes, leading to differences of gene content between subtelomeric and internal regions in *P. falciparum* (Figure S5). For instance, subtelomeric regions in *P. falciparum* are enriched for the rapidly-evolving immune response gene families such as *var, rif, stevor* (Freitas-Junior, et al. 2000a; Kyes, et al. 2007; Hernandez-Rivas, et al. 2013); hence the evolution of these gene families is linked to the mechanisms of karyotype variation.

Given the differences in history of duplicated genes in subtelomeric *versus* internal regions, we evaluate the level of functional constraints/selection in paralogs along chromosomes maps using dN/dS rations (Figures 1.4, S6). We compare patterns for the *var* gene family, which are deployed as the parasite seeks to evade host immune responses (Su, et al. 1995; Scherf, et al. 2008; Claessens, et al. 2014), to paralogs of other gene families in both subtelomeric and internal regions (Figure 1.4). Overall, paralogs of subtelomeric gene families are under less selection constraint than paralogs of internal regions as evidenced by higher dN/dS ratios (Figure 1.4). However, patterns for *var* paralogs seem not affected by their position in the chromosome (Figures 1.4, S6). The varying levels of constraint observed between subtelomeric and internal gene families

suggest that the mechanism of ectopic recombination introduces mutations into gene family members. The more constant level of constraint in the *var* gene family indicates that other forces are at play in diversifying members of this particular gene family, independent of location along chromosome.

1.5.3 Putative origin of each gene of *P. falciparum*

PhyloChromoMap enables exploration of the age and sources of genes along chromosomes. For example, we identify three candidate LGTs (i.e. 1-cys peroxiredoxin, ribosomal protein L35 precursor and holo-ACP synthase, Table S3) as potential EGTs as they encode for apicoplastic functions such as fatty acid synthesis. We can then map these cases of EGT and LGT along chromosomes of *P. falciparum* 3D7 (Figures 1.5, S7). We also bin genes into categories based on possible age (Figure 1.5): LUCA indicates genes in bacteria, archaea and many eukaryotes, LACA are genes only in Archaea and Eukaryotes, and LECA are genes found only among diverse eukaryotes. Importantly, these categorizations should be viewed as putative – they indicate hypotheses and future directions for study.

Table 1.1. Summary of conservation of genes in *P. falciparum*

Description	**Number of occurrences**	
Total in *Plasmodium falciparum* 3D7	5336	
Recent (NIP): In fewer than 10 species in pipeline	3220	(60%)
Older (IP): Phylogenomic pipeline	2116	(40%)
Distribution		
In all major clades of Eukaryotes[a]	1144	(21%)
In at least 4 major clades of Eukaryotes[a]	1440	(27%)
In at least 3 major clades of Eukaryotes[a]	1644	(31%)
In prokaryotes	635	(12%)
In Bacteria and Archaea	267	(5%)
In Bacteria and not in Archaea	202	(4%)
In Archaea and not in Bacteria	166	(3%)

NIP = not in our pipeline, which required ≥10 species to build phylogeny; IP – in pipeline. [a]The five major clades are: SAR (Sr), Archaeplastida (Pl), Opisthokonta (Op), Amoebozoa (Am), and Excavata (Ex). [b]A sequence is considered to be present in a major clade only if it is present on at least 25% of the clades from the next taxonomic rank (e.g. Apicomplexans, Ciliates, Animals, Fungi); sequences in only a few lineages may be contaminants or the result of gene transfers.

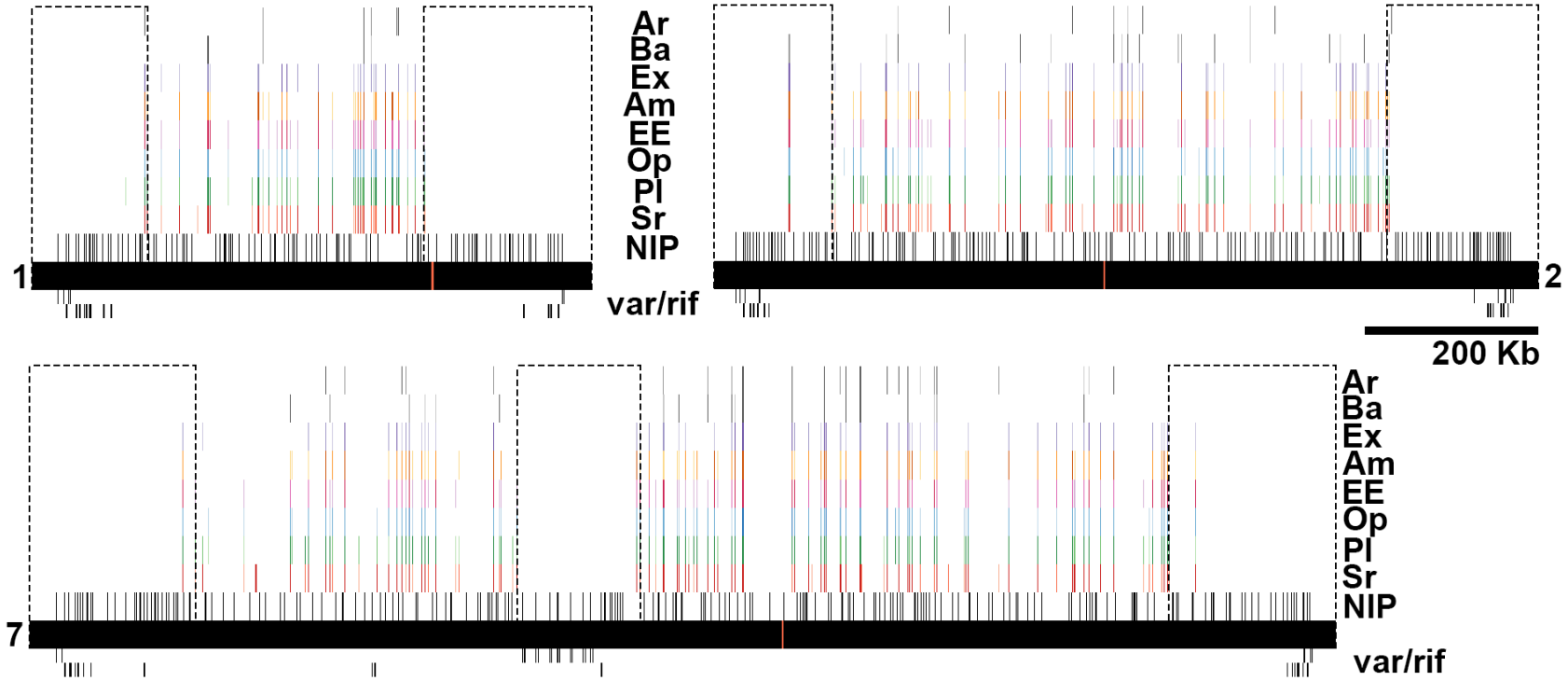

Figure 1.1. Exemplar phylogenomic maps of chromosomes 1, 2 and 7 of *Plasmodium falciparum* 3D7 highlighting 'young' subtelomeric and internal regions (boxes). Black lines represent chromosomes of *P. falciparum* 3D7 and bars above reflect levels of conservation, with dashed boxes around 'young' regions. First row from the bottom (NIP, "not in pipeline") indicates ORFs that do not match our criteria for tree building (i.e. likely *Plasmodium*-specific or mis-annotated ORFs). The remaining rows (bottom to top) are heatmaps reflecting the proportion of lineages of SAR (Sr), Archaeplastida (Pl), Opisthokonta (Op), orphans (EE, "everything else"), Amoebozoa (Am), Excavata (Ex), Bacteria (Ba) and Archaea (Ar) that contain the indicated gene. Shorter lines below the chromosomes show the location of paralogs of *Plasmodium*-specific gene family members involved in antigenic responses: *var* and *rif*.

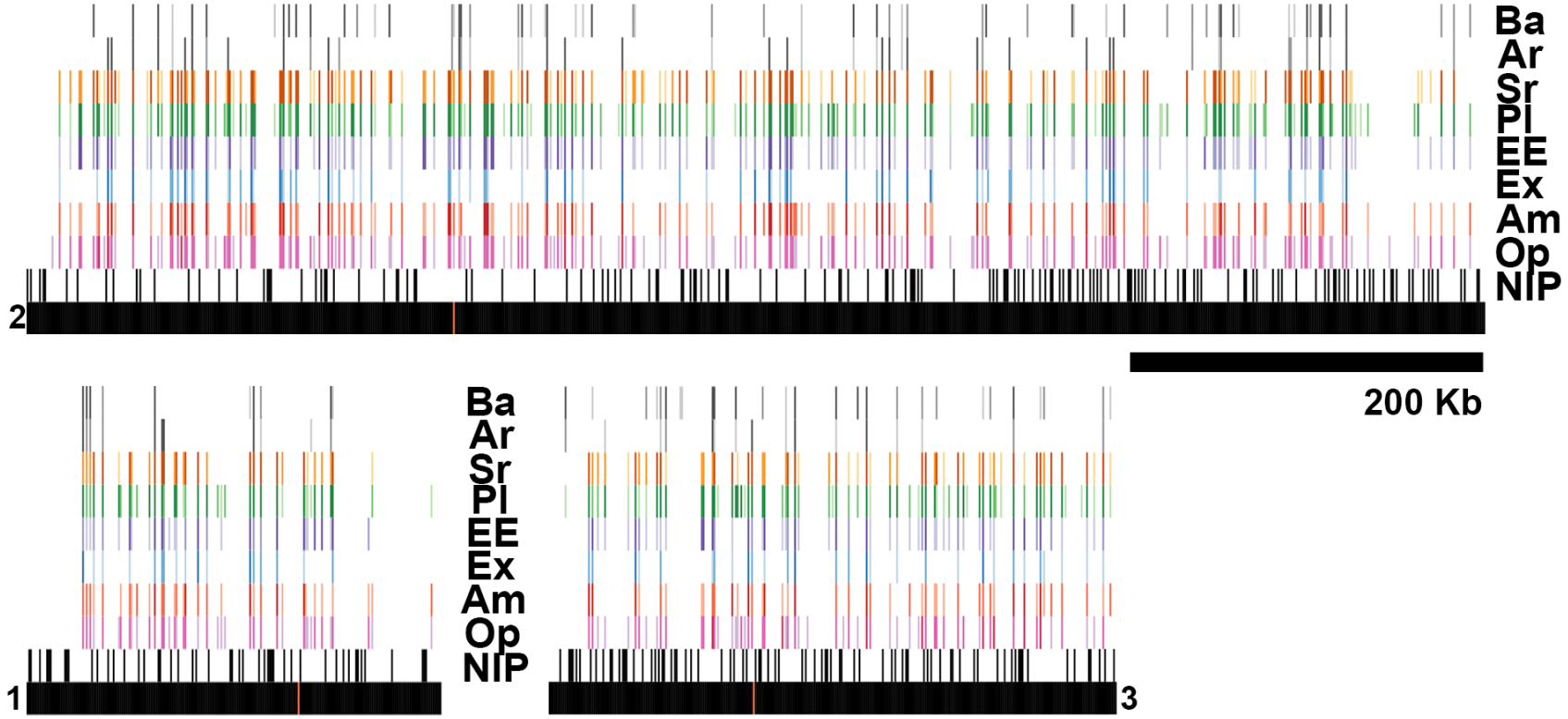

Figure 1.2. Exemplar phylogenomic maps of chromosomes 1-3 of *Saccharomyces cerevisiae* S288C. Black lines represent chromosomes of *S. cerevisiae S288C* and bars above reflect levels of conservation. First row from the bottom (NIP, "not in pipeline") indicates ORFs that do not match our criteria for tree building (i.e. likely *Saccharomyces-specific* or mis-annotated ORFs). The remaining rows (bottom to top) are heatmaps reflecting the proportion of lineages of Opisthokonta (Op), Amoebozoa (Am), Excavata (Ex), orphans (EE, "everything else"), Archaeplastida (Pl), SAR (Sr), Archaea (Ar), Bacteria (Ba) and that contain the indicated gene. Opposite to all the other chromosomes, the chromosome I exhibits large regions of low gene content toward the ends.

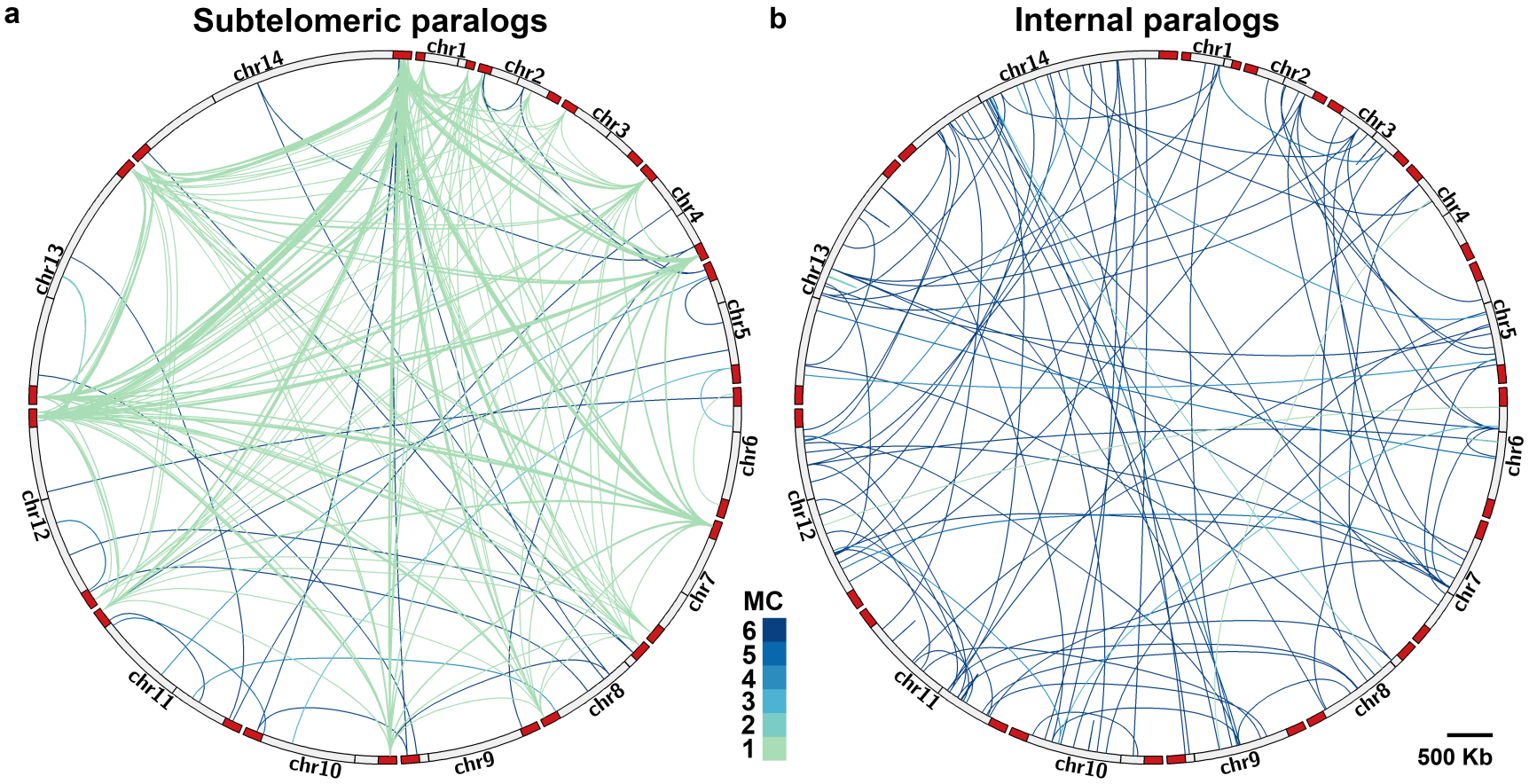

Figure 1.3. Paralogs in a) subtelomeric regions of *P. falciparum* 3D7 tend to be young while paralogs in b) internal regions tend to be old. The 14 chromosomes of *P. falciparum* are displayed as a circle with the red portions of each chromosome indicating subtelomeric regions. The lines within the circles link pairs of paralogs and the color indicates how many eukaryotic major clades (MC, see notes in Figure 1.1) contain those paralogs (i.e. older paralogs are more blue and younger paralogs are more green).

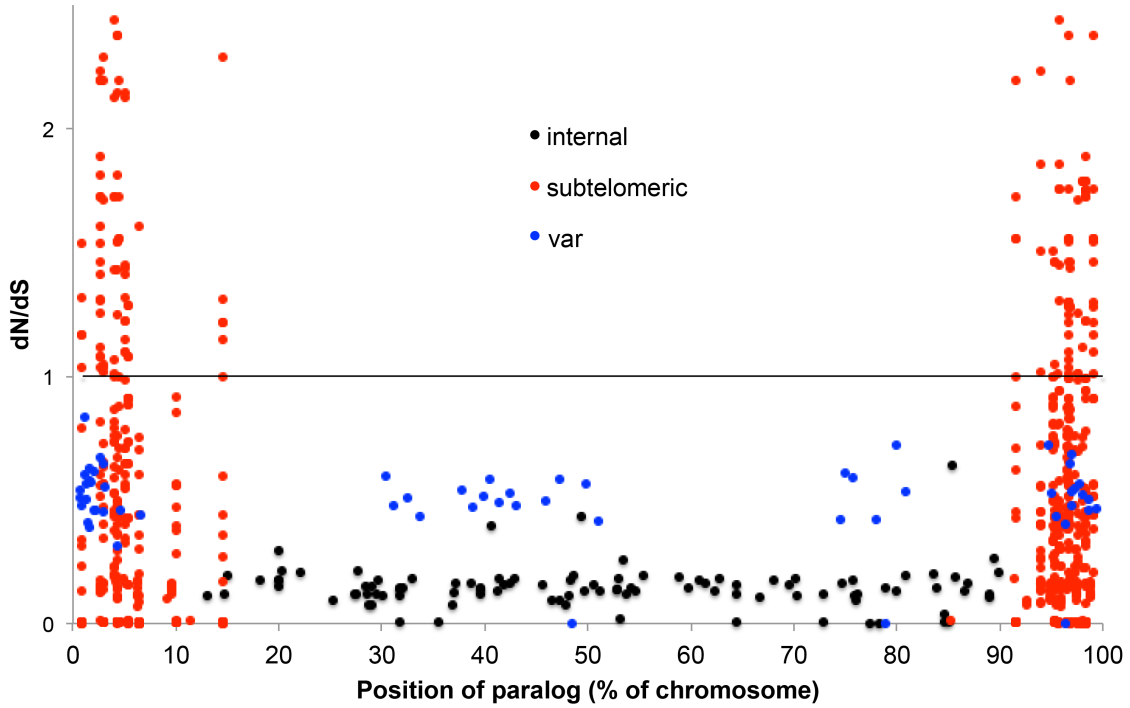

Figure 1.4. Paralogs from gene family *var* (blue) do not exhibit significant differences in selection intensity (i.e. dN/dS) according to location, while paralogs from other gene families (red and black) show significant differences between subtelomeric and internal regions. This graph depicts the dN/dS ratio for three datasets of paralogs, with the x-axis representing the percentage of length of each chromosome, and the graph represents the summary across all 14 chromosomes. Levels of conservation vary among subtelomeric paralogs (red), internal paralogs (black) and paralogs of the gene family *var* (blue). Paralogs exhibit significantly different dN/dS ratios according to their location (Kolmogorov-Smirnov, $p < 0.05$), with subtelomeric paralogs having the highest ranges of dN/dS rations and internal paralogs being under relatively constant levels of constraint. In contrast, dN/dS in *var* paralogs are not affected by location (RELAX, $k = 1.22$, $p >$ 0,05; Figure S6) and are under less functional constraint than most internal paralogs.

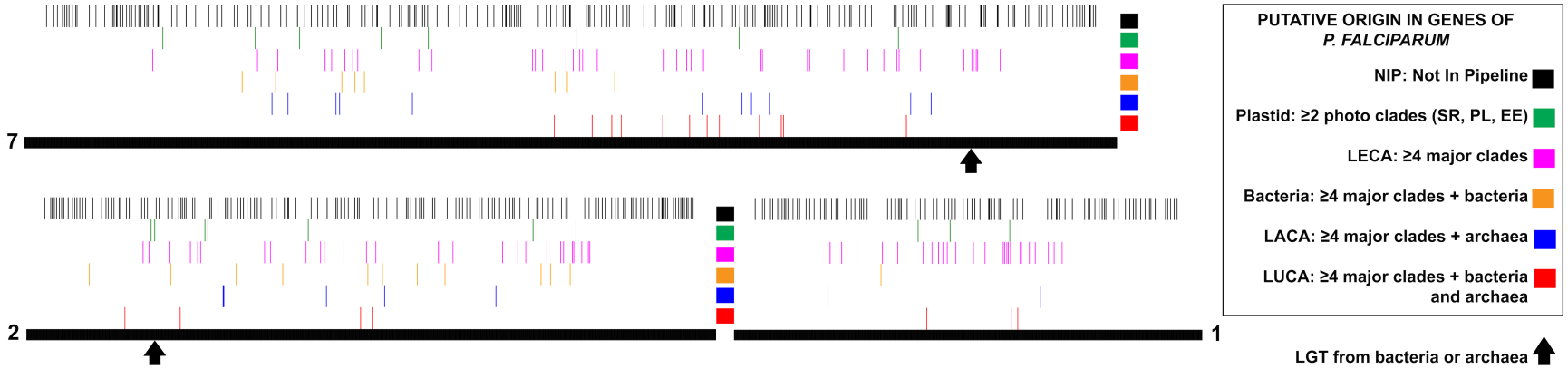

Figure 1.5. Exemplar phylogenomic map of the chromosomes 1, 2 and 7 according to the hypothetical origin of genes. The arrows are candidate LGTs from prokaryotes to Apicomplexa. NIP: not in pipeline, likely young genes, are in black. Candidate EGTs from plastid and mitochondria are in green and orange, respectively. Candidate conserved genes from LECA, LACA and LUCA are in magenta, blue, and red, respectively.

CHAPTER 2

PHYLOTOL: A TAXON/GENE RICH PHYLOGENOMIC PIPELINE TO EXPLORE GENOME EVOLUTION OF DIVERSE EUKARYOTES[2]

2.1 Abstract

Estimating multiple sequence alignments (MSAs) and inferring phylogenies are essential for many aspects of comparative biology. Yet, many bioinformatics tools for such analyses have focused on specific clades, with greatest attention paid to plants, animals and fungi. The rapid increase of high-throughput sequencing (HTS) data from diverse lineages now provides opportunities to estimate evolutionary relationships and gene family evolution across the eukaryotic tree of life. At the same time, these types of data are known to be error-prone (e.g. substitutions, contamination). To address these opportunities and challenges, we have refined a phylogenomic pipeline, now named PhyloToL, to allow easy incorporation of data from HTS studies, to automate production of both MSAs and gene trees, and to identify and remove contaminants. PhyloToL is designed for phylogenomic analyses of diverse lineages across the tree of life (i.e. at scales of >100 million years). We demonstrate the power of PhyloToL by assessing stop codon usage in Ciliophora, identifying contamination in a taxon- and gene-rich database and exploring the evolutionary history of chromosomes in the kinetoplastid parasite *Trypanosoma brucei*, the causative agent of African sleeping sickness. Benchmarking PhyloToL's homology assessment against that of OrthoMCL and a published paper on

[2] Ceron-Romero MA, Maurer-Alcala XX, Grattepanche JD, Yan Y, Fonseca MM, Katz LA. 2019. PhyloToL: A Taxon/Gene-Rich Phylogenomic Pipeline to Explore Genome Evolution of Diverse Eukaryotes. *Mol Biol Evol* 36:1831-1842.

superfamilies of bacterial and eukaryotic organelle outer membrane pore-forming proteins demonstrates the power of our approach for determining gene family membership and inferring gene trees. PhyloToL is highly flexible and allows users to easily explore HTS data, test hypotheses about phylogeny and gene family evolution and combine outputs with third-party tools (e.g. PhyloChromoMap, iGTP).

2.2 Introduction

An important way to study biodiversity is through phylogenomics, which uses the generation of multiple sequence alignments (MSAs), gene trees and species trees (e.g. Katz and Grant 2015; Hug, et al. 2016). During the last two decades, advances in DNA sequencing technology (e.g. 454, Illumina, Nanopore and PacBio) have led to the rapid accumulation of data (transcriptomes and genomes) from diverse lineages across the tree of life, greatly expanding the opportunities for phylogenomic studies (Katz and Grant 2015; Burki, et al. 2016; Brown, et al. 2018; Heiss, et al. 2018). Such approaches are powerful by using increasingly large molecular datasets to reduce the discordance between gene and species trees. Indeed, studies relying on a small number of genes are often impacted by lateral gene transfer, gene duplication and loss, and incomplete lineage sorting (e.g. Maddison 1997; Tremblay-Savard and Swenson 2012; Mallo and Posada 2016). Large-scale phylogenomic analyses allow for the exploration of deep evolutionary relationships (dos Reis, et al. 2012; Wickett, et al. 2014; Katz and Grant 2015; Hug, et al. 2016), but such analyses require data-intensive computing methods. As a result, numerous laboratories have developed custom phylogenomic pipelines proposing different methods to efficiently process and analyze massive gene and taxon databases (e.g. Sanderson, et al. 2008; Wu and Eisen 2008; Smith, et al. 2009; Kumar, et al. 2015).

In general, phylogenomic pipelines are composed of three steps: 1) construction of a collection of homologous gene datasets from various input sources (e.g. whole genome sequencing, transcriptome analyses, PCR based studies), 2) production of MSAs, and 3) generation of gene trees and sometimes a species tree. Phylogenomic pipelines typically put more effort in the first two steps (collecting homologous genes and MSA curation) to ensure a more accurate tree inference. For instance, pipelines such as PhyLoTA (Sanderson, et al. 2008) and BIR (Kumar, et al. 2015) focus on the identification and collection of homologous genes by exploring public databases such as GenBank (Benson, et al. 2017). On the other hand, pipelines such as AMPHORA (Wu and Eisen 2008) and Mega-phylogeny (Smith, et al. 2009) focus on the construction and refinement of robust alignments rather than the collection of homologs. A recently published tool, SUPERSMART (Antonelli, et al. 2017), incorporates more efficient methods for data mining than PhyLoTA (Sanderson, et al. 2008). SUPERSMART includes sophisticated methods for tree inference using a multilocus coalescent model, which benefits biogeographical analyses. Although these pipelines incorporate sophisticated methods for data mining, alignment and tree inference, a major issue is that they are optimized for either a relatively narrow taxonomic sampling (e.g. plants) or for relatively narrow sets of conserved genes/gene markers.

A major problem for phylogenomic analyses using public sequence data, including GenBank and EMBL (Baker, et al. 2000), is the inherent difficulty in identifying and removing annotation errors and contamination (e.g. data from food sources, symbionts or organelles). Additional errors are introduced when non-protein coding regions (e.g. pseudogenes, promoters and repeats) are inferred as open reading

frames (ORFs) by gene-prediction tools such as GENESCAN (Burge and Karlin 1997), SNAP (Korf 2004), AUGUSTUS (Stanke and Morgenstern 2005) and MAKER (Cantarel, et al. 2008). Similarly, some public databases are more prone to contain annotation errors than others depending on how much effort they invest in manual curation of public submissions. For instance, data from GenBank NR, TrEMBL (Bairoch and Apweiler 2000) and KEGG (Kanehisa and Goto 2000) may have very high rates of these errors, whereas curated resources like Gene Ontology (GO; Ashburner, et al. 2000) and SwissProt (Bairoch and Apweiler 2000) are more likely to have low to moderate rates of such errors (Schnoes, et al. 2009). The misidentification errors in these databases often stem from problems surrounding accurate taxonomic identification of sequences from HTS data sets, as contamination by other taxa can be frequent, particularly of organisms that cannot be cultured axenically (Shrestha, et al. 2013; Lusk 2014; Parks, et al. 2015). Hence, a crucial element of any phylogenomic pipeline that relies on public databases is the ability to identify and exclude annotation errors and contaminants from its analyses.

At the same time, the availability of curated databases and third-party tools provide considerable power and efficiency for phylogenomic analyses. We rely on OrthoMCL, a database generated initially to support analyses of the genome of *Plasmodium falciparum* and other apicomplexan parasites (Li, et al. 2003; Chen, et al. 2006), for the initial identification of homologous gene families (i.e. GFs). We also incorporate GUIDANCE V2.02 (Penn, et al. 2010; Sela, et al. 2015b) for assigning statistical confidence MSA scores based on the robustness of the MSA to guide-tree uncertainty. GUIDANCE allows an efficient identification and removal of potentially

non-homologous sequences (i.e. sequences having very low scoring values) and unreliably aligned columns and residues under various parameters (Privman, et al. 2012; Hall 2013; Vasilakis, et al. 2013). This flexibility is critical – while concepts such as homology and paralogy have clear definitions in textbooks, when it comes to deploy phylogenomic tools on inferences at the scale of >100 million years, they become working definitions that depend of parameters and sampling of both genes and taxa. Finally, we have chosen RAxML V8 (Stamatakis, et al. 2005; Stamatakis 2014) for tree inference as its efficient algorithms allow for robust estimation of maximum likelihood trees [though users can access the MSAs from our pipeline for analyses with other software].

Our original phylogenomic pipeline aimed to explore the eukaryotic tree of life using multigene sequences available in GenBank from diverse taxa (Grant and Katz 2014a; Katz and Grant 2015). This first version generated a collection of ~13,000 gene families (i.e. GFs) from ~800 species distributed among Eukaryota, Bacteria and Archaea, and included a suite of methods to process gene alignments and trees. The 800 species were a subset of available taxa, picked to represent, more or less evenly, the main eukaryotic lineages with no more than two species per genus. Moreover, although the focus was on eukaryotes, bacteria and archaea were also included in order to allow detection of contamination, lateral gene transfer events and/or for exploring phylogenetic relationships that include all cellular life. GFs originally defined by OrthoMCL were used as seeds to search more homologous sequences from additional taxa. Then, the enriched GFs pass for an additional quality-check step that re-evaluates homology. This step includes applying a combination of methods that include removing alleles and

nonhomologous genes and highly-divergent sequences based on pairwise comparisons with Needle (Rice, et al. 2000), with robust alignments produced with MAFFT (Katoh and Standley 2013) that were then filtered with GUIDANCE. These refined high-quality MSAs were used to produce gene trees with RAxML. An additional option is to identify orthologs based on their position in gene trees, which can be used to generate concatenated alignments for species tree inference (see Grant and Katz 2014a for more details).

This new version, which we name PhyloToL (Phylogenomic Tree of Life), incorporates significant improvements over Grant and Katz (2014a), including a more efficient method to capture HTS data, a more robust homology detection approach, a novel tree-based method for contamination removal, and substantially more efficient scripts and improved databases. PhyloToL contains a database of 13,103 GFs that include up to 627 eukaryotes (58 generated in our lab), 312 bacteria and 128 archaea. Here we describe our updated approaches providing examples of stop codon usage assessment in Ciliophora and detection of contamination produced by many HTS studies (including our own). We also illustrate the potential of PhyloToL by depicting the evolutionary history of the genes on the chromosomes of the human parasite *Trypanosoma brucei*, causative agent of African sleeping sickness.

2.3 New approaches

PhyloToL (https://github.com/Katzlab/PhyloTOL; last updates January 2019) is divided in four major components: 1) Gene family assessment per taxon, 2) refinement of homologs and gene tree reconstruction, 3) tree-based contamination removal and 4)

generation of a supermatrix for species tree inference (i.e. concatenation). The first component starts with data from either public databases or those generated by our own 'omics projects and categorizes sequences into a collection of candidate GFs. This part of PhyloToL includes steps for removing bacterial contamination (given our focus on eukaryotes) and translating sequences using the most appropriate inferred genetic code (Figure 2.1*A*). The second component includes a series of steps to assess homology in the candidate GFs based on sequence similarity, sequence overlap, and refinement of MSAs prior to reconstructing phylogenies (Figure 2.1*B*). The third component includes a novel method that iterates the second component (refinement of homologs and gene tree reconstruction) to remove contamination inferred from phylogenetic trees (Figure 2.1*C*), which is critical given the high frequency of contamination in many HTS datasets. While the combination of methods in the first three components identify homologs within GFs (see MATERIALS AND METHODS), the distinction between paralogous and orthologous sequences occurs only in the optional fourth component. This component detects orthologous sequences based on their position in phylogenetic trees and concatenates them into a supermatrix for species tree inference (Figure 2.1*D*); this last component has not been modified since the last published version of the pipeline (Grant and Katz 2014a; Grant and Katz 2014b; Katz and Grant 2015), and users can explore other tools for concatenation (Leigh, et al. 2008; Narechania, et al. 2012; Drori, et al. 2018; Vinuesa, et al. 2018) using the single gene MSAs generated by PhyloToL.

Additional to the primary goal of PhyloToL, which was reconstructing the evolutionary history of eukaryotes, this new version emphasizes the flexibility to allow studies of GFs evolution as well as phylogenomics with varying parameters and

taxon/gene inclusion. Though there are many other tools out there for phylogenomic analyses (e.g. OneTwoTree (Drori, et al. 2018), SUPERSMART (Antonelli, et al. 2017) and PhyloTA (Sanderson, et al. 2008)), we believe PhyloToL is distinctive because of its combination of: 1) inclusion of both database and user-inputted data; 2) focus on broad taxon inclusion for 'deep' events (e.g. ≥100 million years); and 3) flexibility for exploration of multiple hypotheses and parameters (Table S4).

2.3 Results and discussion

The overall structure of PhyloToL was improved over Grant and Katz (2014a) by dividing the pipeline into 4 major components (Figure 2.1) allowing different modes to execute these components depending on the type of study. PhyloToL also includes new methods to use data from more sources (in component 1, Figure 2.1*A*), refine MSAs from GFs (in component 2, Figure 2.1*B*), and to remove contaminant sequences (in component 3, Figure 2.1*C*). Here we explain improvements on the overall structure of PhyloToL and benchmark the performance of new methods by analyses of ancient gene families.

2.3.1 Pipeline structure

Although PhyloToL is designed for phylogenomic analyses of diverse lineages across the tree of life, it can also be deployed in different ways for a variety of purposes such as phylogenomic chromosome mapping (Cerón-Romero, et al. 2018), gene discovery, or metatranscriptomics. For instance, the GF assessment per taxon, refinement of GFs and gene tree reconstruction (i.e. first and second components of PhyloToL) can be run independently, and the tree-based contamination removal and generation of a supermatrix (third and fourth components) are optional. Moreover, the user can also run

the second component in two alternative modes: i) only quality control (QC) for GFs and ii) without gene tree. Running the second component of PhyloToL only for QC for GFs is helpful when the primary aim is to collect sequences for candidate GFs (QC involves filtering sequences by length, overlap and similarity, see MATERIALS AND METHODS) or for exploring taxonomic diversity within each gene family. Likewise, running the second component of PhyloToL without generating gene trees is useful for inspecting regions of homology (motif searching), trying alternative methodologies (i.e. those other than RAxML V8, which is incorporated into PhyloToL) for phylogenetic tree inference and to simply create a curated database of aligned homologous proteins (i.e. having sequences with divergence levels above the defined threshold removed by GUIDANCE). Our approach for determining homology is through generation of MSAs using GUIDANCE V2.02 (Penn, et al. 2010; Sela, et al. 2015b) with sequence and column cutoff 0.3 and 0.4, respectively, to determine which sequences meet criteria for retention. These GUIDANCE parameters were chosen based on inspection of early runs of our data because the default parameters in GUIDANCE are geared for shallower levels of diversity and tend to exclude much of our focal taxa. Indeed, GUIDANCE scores are alignment dependent and so cutoffs are empirically defined. As described in our manual (https://github.com/Katzlab/PhyloChromoMap_py/blob/master/phylochromomap_manual.pdf) users can change these parameters for their own data sets in order to explore homology more deeply.

2.3.2 Performance of PhyloToL in GF estimation per taxon

To exemplify outputs of the first component of PhyloToL, GF assessment per taxon, we provide data from RNA-seq studies of the ciliates *Blepharisma japonicum*

(MMETSP1395) and *Strombidium rassoulzadegani* (MMETSP0449_2). Each of these two datasets starts with > 20,000 assembled transcripts, from which ~1% are contamination from rRNAs, bacterial and archaeal sequences that are removed (Table 2.1). The final datasets after running through PhyloToL (only the GF assessment per taxon component) contain between 5,000 and 10,000 transcripts assigned to eukaryotic GFs and representing ~20% of the initial set of sequences (Table 2.1). PhyloToL also allows us to assess that *B. japonicum* potentially uses the "*Blepharisma*" genetic code (i.e. UAR as stop codon, UGA is translated to tryptophan; Lozupone, et al. 2001; Sugiura, et al. 2012) and *S. rassoulzadegani* uses the "ciliate" genetic code (i.e. only use UGA as stop codon, and UAR is reassigned to glutamine; Caron and Meyer 1985).

We evaluated the importance of PhyloToL's inspection of putative stop codons for these two taxa by also processing the transcriptomic data forcing translation with the universal and the "ciliate" genetic codes (Figure 2.2*A*). Here we found that when using PhyloToL's inferred alternative genetic code, transcripts were substantially longer than when forced to be processed with universal or ciliate genetic codes (Figure 2.2*A*), which suggests that using the carefully assessed genetic code allows the user to retrieve a larger proportion of each transcript.

2.3.3 Performance of PhyloToL in tree-based contamination removal

We then tested the third component of PhyloToL (i.e. tree-based contamination removal) using a dataset of 152 GFs that includes up to 167 taxa distributed among eukaryotes, bacteria and archaea. To give the user a sense of the time involved, using a computer with 128 GB of RAM and 10 cores, the analyses took 86 hours and 5 iterations

of contamination removal. However, 79% of the contaminant sequences were removed in the first iteration, which also took 52% of the total time (Figure 2.2*B*).

Contaminant sequences detected often originated from food sources or endosymbiosis (at least 52% and 42% of the total contaminats, respectively). For instance, sequences from the amoeba *Neoparamoeba* are often nested within Euglenozoa (in 14 GFs; Figure 2.3*A*) because likely some of its data are actually from a (past or present) kinetoplastid endosymbiont as previously reported by Tanifuji et al. (2011). Likewise, sequences from the foraminifera *Sorites*, which hosts a dinoflagellate endosymbiont (Langer and Lipps 1995), are sometimes nested within dinoflagellate sequences (37 GFs; Figure 2.3*B*). On the other hand, sequences from the Katablepharid *Roombia truncata* are sometimes nested among the SAR clade as sister to Stramenopila (in 3 GFs; Figure 2.3*C*); these sequences are potentially from diatoms, which are used for feeding *R. truncata* (Okamoto, et al. 2009). Finally, sequences from the Rhizaria *Leptophrys vorax,* which is fed on green algae, are often nested among green algal clades (38 GFs; Figure 2.3*D*).

Using the methods developed here, users can identify sources of contamination in individual taxa and then remove contaminating sequences in PhyloToL's contamination loop. This step is critical because sequence contamination is a common problem in HTS data of public databases (Merchant, et al. 2014; Kryukov and Imanishi 2016). Indeed, previous studies have demonstrated that sequence contamination is one of the most important obstacles for evolutionary studies (Laurin-Lemay, et al. 2012; Struck 2013; Philippe, et al. 2017).

2.3.4 Implementation for phylogenomic chromosome mapping

To exemplify an implementation of PhyloToL, we combined outputs with our tool PhyloChromoMap (Cerón-Romero, et al. 2018) to explore the evolutionary history of chromosomes in the kinetoplastid parasite that causes African sleeping sickness, *Trypanosoma brucei gambiense* DAL972 (assembly ASM21029v1). Combining these tools, with PhyloChromoMap for mapping genes along each strand separately, we generated a map that displays the evolutionary history of 9,755 genes across both strands of the *T. brucei gambiense* chromosomes (Figures 2.4, S8).

Previous studies have shown that karyotypes of kinetoplastid parasites have large syntenic polycistronic gene clusters (PGC), where genes are sequentially arranged on the same strand of DNA and expressed as multi-gene transcripts (Berriman, et al. 2005; El-Sayed, et al. 2005; Daniels, et al. 2010; Martinez-Calvillo, et al. 2010). We observed that almost all genes matching our GFs fall in PGCs and have a wide distribution throughout all 11 chromosomes, with variable gene density among chromosomes (Figures 2.4, S8). Besides the presence of PGCs in *T. brucei*, previous studies proposed that large subtelomeric arrays of species-specific genes might serve as breakpoints for ectopic recombination in the nuclear membrane (Berriman, et al. 2005; El-Sayed, et al. 2005), a phenomenon that is also described in the apicomplexan parasite, *Plasmodium falciparum* (Freitas-Junior, et al. 2000b; Scherf, et al. 2001; Hernandez-Rivas, et al. 2013; Cerón-Romero, et al. 2018). However, while young and highly recombinant subtelomeric regions of at least 58 Mbp (up to 218 Mbp) are present in all *P. falciparum* chromosomes (Cerón-Romero, et al. 2018), in *T. brucei gambiense* this pattern is only evident in chromosomes 3 and 9 (Figure S8). This indicates that although ectopic recombination of

subtelomeric regions can play a role in the karyotype evolution of *T. brucei*, it may not be as crucial to the success of this parasite as compared to *P. falciparum*.

We also explored the level of evolutionary conservation of genes in *T. brucei gambiense* based on their phylogenetic distribution as estimated by PhyloToL. Here, we detected that genes tend to be either very conserved or very divergent, with few genes of intermediate conservation (χ^2, $p < 0.05$; Figure S9). About 73% of the published genes in the *Trypanosoma brucei gambiense* DAL972 (assembly ASM21029v1) genome lacked homologs to any of our GFs and thus may be *Trypanosoma*-specific genes and/or mis-annotations (Table 2.2). Of the remaining 27% of genes that match conserved eukaryotic GFs, ~44% are conserved among all the major eukaryotic clades, ~8% are shared between all major eukaryotic clades and Archaea and ~8% are conserved among all major eukaryotic clades, Archaea and Bacteria (Table 2.2).

2.3.5 Test of homology assessment

To benchmark the homology assessment in PhyloToL, we compared reconstructions of ancient (i.e. present in bacteria, archaea and eukaryotes) gene families originally estimated in OrthoMCL. Members of ancient gene families tend to be categorized in different orthologous groups in OrthoMCL (e.g. α-tubulin is group OG5_126605 and β-tubulin is group OG5_132171). We analyzed 8 ancient gene families that were likely present in LUCA: ATPases, family B DNA polymerase, elongation factors Tu/1a, elongation factors G/2, glutamyl- and glutaminyl-tRNA synthetases, RNA polymerase subunit A, RNA polymerase subunit B and tubulins. Overall, our recovery of the homology of these ancient GFs was robust to our taxon-rich analyses (Figure 2.5).

For four of the eight gene families (i.e. glutaminyl-tRNA synthetases, RNA polymerase subunit A, RNA polymerase subunit B and tubulins) there were a few cases (<0.05%) where sequences were misclassified in the earlier steps of PhyloToL, likely due to the limited taxon sampling in the OrthoMCL-based 'seeds' for BLAST analyses.

We also benchmarked PhyloToL against the reconstruction of gene families of bacterial and eukaryotic organelle outer membrane pore-forming proteins as proposed by Reddy and Saier (2016). Reddy and Saier (2016) combined 76 gene families among 5 superfamilies of varying size. To compare their homology statements to inferences from PhyloToL, we focused on the 12 gene families already included in the PhyloToL databases that fall into two superfamilies, the prokaryotic superfamily I (SFI) and eukaryotic superfamily IV (SFIV). Under PhyloToL's default parameters (i.e. GUIDANCE V2.02 sequence cutoff = 0.3, column cutoff = 0.4, number of iterations = 5), many SFI members (different GFs) determined by Reddy and Saier (2016) do not meet our criteria for homology: when running the full set of sequences of SFI in PhyloToL, only sequences of the largest GF survive, indicating that the other GFs are too dissimilar to be included in a MSA under our parameters (Table S5). We then re-ran PhyloToL to test homology in every cluster and sub-cluster of GFs that form SFI but at the end only cluster III meets our conservative criteria for homology (Figure 2.6, Table S4). In contrast to SFI, both members of the eukaryotic SFIV are retained under default parameters in PhyloToL (Figure 2.6, Table S5). We then forced the gene families determined by Reddy and Saier (2016) to align, and found limited evidence of homology (e.g. conserved columns in MSAs). In sum, our estimation of homology is more stringent

than in Reddy and Saier (2016), and the exploration of this question took ~3 hours on a computer with 4 threads, highlighting the flexibility of PhyloToL for users.

2.4 Materials and methods

There are four components in PhyloToL's algorithm: 1) GF assessment per taxon, 2) refinement of GFs and gene tree reconstruction, 3) tree-based contamination removal and 4) generation of a supermatrix for species tree inference. The GF assessment per taxon includes features such as translation using informed genetic codes. The refinement of GFs and gene tree reconstruction filters and asserts homology in the GFs comparing sequences by length, overlap, similarity and MSA. The component tree-based contamination removal detects and removes contaminant sequences based on predefined contamination rules and the position of the sequences in gene trees. Finally, the component generating a supermatrix for species tree inference chooses orthologs and discards paralogs based on tree topology in order to concatenate MSAs for species tree inference.

2.4.1 Naming sequences

PhyloToL uses standardized names that are compatible with the third-party tools incorporated into the pipeline (e.g. GUIDANCE, RAxML). Although the users are free to assign different codes to the taxa at their convenience, PhyloToL requires that every taxon is named using a 10-digit code that broadly reflects its taxonomy; this code is divided in three components, a major clade (e.g. Op = Opisthokonta), a "minor" clade (e.g. Op_me = Metazoa) and a species name (e.g. Op_me_hsap for Homo sapiens). For each sequence, the 10 digit-code is followed by the sequence identifier such as the

GenBank accession or Ensembl ID (e.g. Op_me_hsap_ENSP00000380524). This naming system allows an easy control of names when handling alignments and trees.

2.4.2 GF assessment per taxon

The first component of PhyloToL (i.e. GF assessment per taxon; Figure 2.1*A*) allows the inclusion of a large number of data sources from online repositories (e.g. GenBank) or from the user's lab, and of different types (e.g. transcriptomes, proteins or annotated proteins from genomic sequences (e.g. 454, Illumina, ESTs)). The first steps aim to accurately assign sequences to homologous GFs, with improvements to the efficiency of these processes as compared to our original pipeline (Grant and Katz 2014a; Grant and Katz 2014b; Katz and Grant 2015). To exemplify methods, we focus on the inclusion of Illumina transcriptome data, though the structure can easily be adapted for other sources. PhyloToL uses a pipeline for passing assembled transcripts through a variety of steps for: removal of short contigs (at a user-defined length), removal of putative contaminants (from ribosomal RNAs (rRNA), bacteria and archaea), and assess gene families. To remove rRNA sequences, we rely on BLAST, comparing each sequence against a database of diverse rRNA sequences sampled from across the tree of life (75 bacteria, 26 archaea and 77 eukaryotes). This is followed by the identification and removal of bacterial/archaeal transcripts through USEARCH V10 (Edgar 2010), which compares data against both a database of diverse bacterial + archaeal proteins and another database of diverse eukaryotic proteins, retaining all non-bacterial/archaeal transcripts (i.e. those with strong matches to eukaryotes, and those remaining unassigned). With this pruned dataset, USEARCH is again used to bin these eukaryotic-enriched sequences into

OrthoMCL GFs while rRNA and bacterial/archaeal transcripts are saved in a different location for easy retrieval if desired.

With growing evidence for the diversity of stop codon reassignments across the eukaryotic tree of life (Keeling and Doolittle 1997; Lozupone, et al. 2001; Keeling and Leander 2003; Heaphy, et al. 2016; Swart, et al. 2016; Panek, et al. 2017), we include an optional step to evaluate potential alternatives to conventional stop codon usage (frequent in frame non-conventional stop codons). This step is essential for some clades such as Ciliophora, where there are at least eight unconventional genetic codes (i.e. not all three traditional stop codons terminate translation). Using the most appropriate genetic code, each nucleotide sequence is then translated into the corresponding amino acid ORF.

Given the imperfect nature of HTS data, we take a conservative approach to avoid inflating the number of paralogs for each taxon and, therefore, we remove nearly identical sequences. These nearly identical sequences can represent an unknown mixture of alleles, recent paralogs and more importantly sequencing and/or assembly errors, which can be problematic for the comparative aspects of PhyloToL. To avoid this issue, for every taxon we remove nearly identical sequences at the nucleotide level (> 98% nucleotide identity across ≥ 70% of their length).

An additional step is available to address the well-known phenomenon of sample bleeding (also known as index switching; Mitra, et al. 2015; Larsson, et al. 2018) that occurs during Illumina sequencing. Based on the observation that some of our taxa were contaminated by one another during Illumina sequencing, we developed a method to remove low read coverage contigs that are identical to higher read coverage contigs. To

this end, we performed a USEARCH ("BLAST") all vs. all of the nucleotide ORFs (at a minimum identity of 98% across ≥ 70% of their length). Those sequences that form clusters of hits to other taxa represent potential cross-contaminants. Next, those sequences with a substantially high read coverage compared to the mean (e.g. 10x more than the mean) are retained and low-read coverage sequences as excluded. In ambiguous cases (i.e. all are low read number), the entire group of sequences is discarded. Although this step is highly dependent on transcriptional state and sequencing depth, this conservative approach impacts < 5% of transcripts for a given taxon using our own Illumina data.

2.4.3 Refinement of homologs and gene tree reconstruction

In the second component of PhyloToL (i.e. refinement of homologs and gene tree reconstruction; Figure 2.1*B*), GFs pass through a procedure to assess homology and then to produce gene trees. The procedure starts with a QC step that includes two filters: an overlap filter and a similarity filter. The overlap filter aims to remove non-homologous sequences, which are sequences substantially longer than putative homologs (e.g. those with only shared motifs), or atypically short (i.e. those with insufficient overlap). Such sequences will confound paralog counting and can negatively impact the alignments. To proceed, we start by identifying a 'master sequence' as the putative homolog. This sequence has the lowest E-value from the GF assignment and is also ≤150% the average length of the members from the reference GF dataset. We then retain all sequences that have a pairwise local alignment overlap that includes at least 35% of the length of the master sequence. In contrast, the optional similarity filter allows the user to remove alleles and recent paralogs (i.e. too similar sequences) at a user-defined cutoff to improve

efficiency. The similarity filter uses an iterative process in which the next longest sequence acts as the 'master sequence' to remove highly similar sequences, and repeats until there are no more sequences that can be assigned as a 'master sequence'.

For the next part of the procedure to assess homology within each GF, PhyloToL relies on GUIDANCE V2.02 scores, and using a user-specified number of iterations, identifies and removes unreliably aligned and potentially non-homologous sequences (Figure 2.1*B*). Then, GUIDANCE is used to filter the final alignment using preset cutoffs for sequences and columns (default parameters or empirically defined, in our case 0.3 for sequences and 0.4 for columns). In contrast to the previous version of the pipeline that relied on only two iterations of GUIDANCE, one for removing poorly-aligned sequences and another for removing poorly-aligned columns, PhyloToL iterates the sequence-removal step either for a user-defined number of iterations or until all unreliable sequences have been removed. Only then the columns are removed based on the user-specified confidence threshold score (the default number of bootstrap replicates for each GUIDANCE run is 10). Residues with low confidence scores, based on a settable residue score cutoff, can be masked in the alignment with an "X" (turned off in our defaults). Finally, in PhyloToL, GUIDANCE uses more accurate MAFFT V7 parameters, including an iterative refinement method (E-INS-i algorithm, and up to 1000 iterations). The E-INS-i algorithm was chosen because it makes the smallest number of assumptions of the three iterative refinement methods implemented in MAFFT and is recommended if the nature of sequences is less clear.

2.4.4 Tree-based contamination removal

The third component of PhyloToL (i.e. tree-based contamination removal; Figure 2.1*C*) includes a method to identify and remove contaminants based on their location within the phylogenetic trees, though user scrutiny of results is required. If inspection of gene trees reveals sequences from a given taxon frequently nested among distantly related lineages, the user can create a set of "rules for contamination removal" and then run the tree-based contamination removal that will detect and remove potential contaminants from the alignments and subsequent trees (Figure 2.1*C*). To help users to define their rules for contamination removal, PhyloToL also generates a report (summary_contamination.csv) containing the frequency of every sister clade per lineage ignoring those with significantly longer branches than the average branch length of the tree, which allows the users to differentiate contamination (e.g. food, symbionts and other sources) from fast evolving taxa that were incorrectly placed in trees. This component of PhyloToL iterates the refinement of homologs and gene tree reconstruction (i.e. second component) using the pre-defined rules to identify sequences of contamination and removing them for the next iteration. This continues until no more 'contaminant' sequences are identified. The component tree-based contamination removal also produces a full list of contaminant sequences that can be removed from the permanent databases. In order to run the tree-based contamination removal more efficiently, potentially non-homologues (i.e. sequences discarded by GUIDANCE) are also removed in every iteration.

Table 2.1. Summary of the experiment of gene family assessment per taxon.

Sequences	***Blepharisma japonicum***	***Strombidium rassoulzadegani***
Original assembly	45,231	24,810
Removed rRNA	114	33
Removed prokaryotic	453	290
Assigned to PhyloToL GF	10,060	4,764

Table 2.2. Summary of conservation of genes in *Trypanosoma brucei*.

Description	Number of genes[b]
Total in *Trypanosoma brucei*.	9755
Recent (NIP): Not in PhyloToL[a]	7125
Older (IP): In PhyloToL[a]	2630
Distribution	
Only in eukaryotes	
1 major clade	39
2 major clades	85
3 major clades	113
4 major clades	190
5 major clades	385
All major clades (including EE)	1150
In eukaryotes and prokaryotes	
Eukarya, Archaea and Bacteria[c]	205
Eukarya and Archaea[c]	207
Eukarya and Bacteria[c]	185
Excavata and either Bacteria or Archaea	2

[a] NIP = did not meet the requirement of ≥ 4 sequences (from the 167 taxa that were chosen for this study) to produce a tree, and are therefore likely either very divergent or misannotated. [b] A gene is considered to be present in a major clade only if it is present in at least 25% of the clades from the next taxonomic rank (e.g. Euglenozoa in Excavata, Apicomplexa in SAR, Animals or Fungi in Opisthokonta); sequences in only a few lineages may be contaminants or the result of gene transfers. [c] In at least 5 eukaryotic major clades: Excavata (Ex), Archaeplastida (Pl), SAR (Sr), Amoebozoa (Am) and Opisthokonta (Op). For every tree the root was placed in between Bacteria and Archaea + Eukaryotes when there were Bacteria; between Archaea and Eukaryotes when there were not Bacteria; or in Opisthokonta when there were not prokaryotes (Katz and Grant 2015).

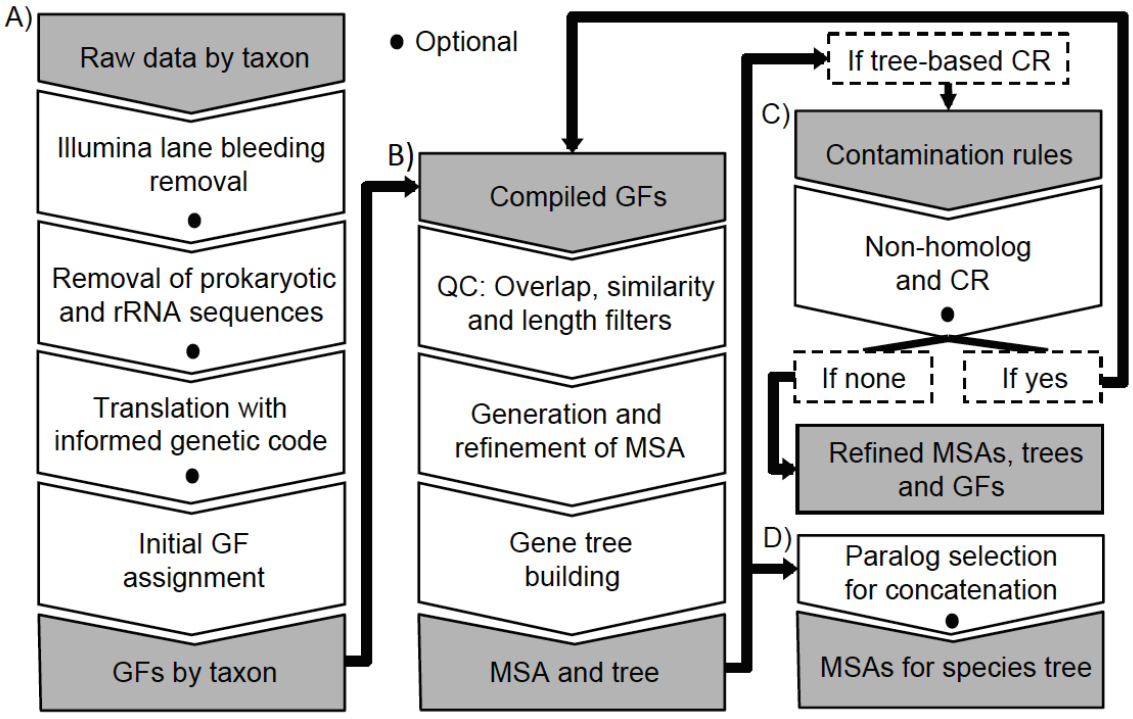

Figure 2.1. The four components of PhyloToL. GF = Gene Family, QC = Quality Control, CR = Contamination Removal. A) The first component processes and classifies raw data from different sources (e.g. transcriptomes, genomes, and protein data) into a collection of gene families. In the initial step, transcriptomes produced in-lab are processed to identify and remove sample bleeding (Mitra, et al. 2015) in an Illumina lane (cross-contamination). Then, prokaryotic sequences and rRNA sequences are removed from transcriptomes. Finally, transcriptomic and genomic sequences are translated using informed genetic codes. B) The second component compiles all gene families by taxon in the gene family database, refines an MSA, and produces a phylogenetic tree for each gene family. C) The third component (optional) detects contaminant sequences using gene trees and pre-defined contamination rules, and also detects non-homologous sequences after the MSA refinement process. Contaminants and non-homologs are identified and removed from the gene family database iteratively. D) The fourth component (optional) identifies orthologous sequences using a tree-based approach for removing paralogs. Alignments of orthologs can be concatenated to produce a species tree.

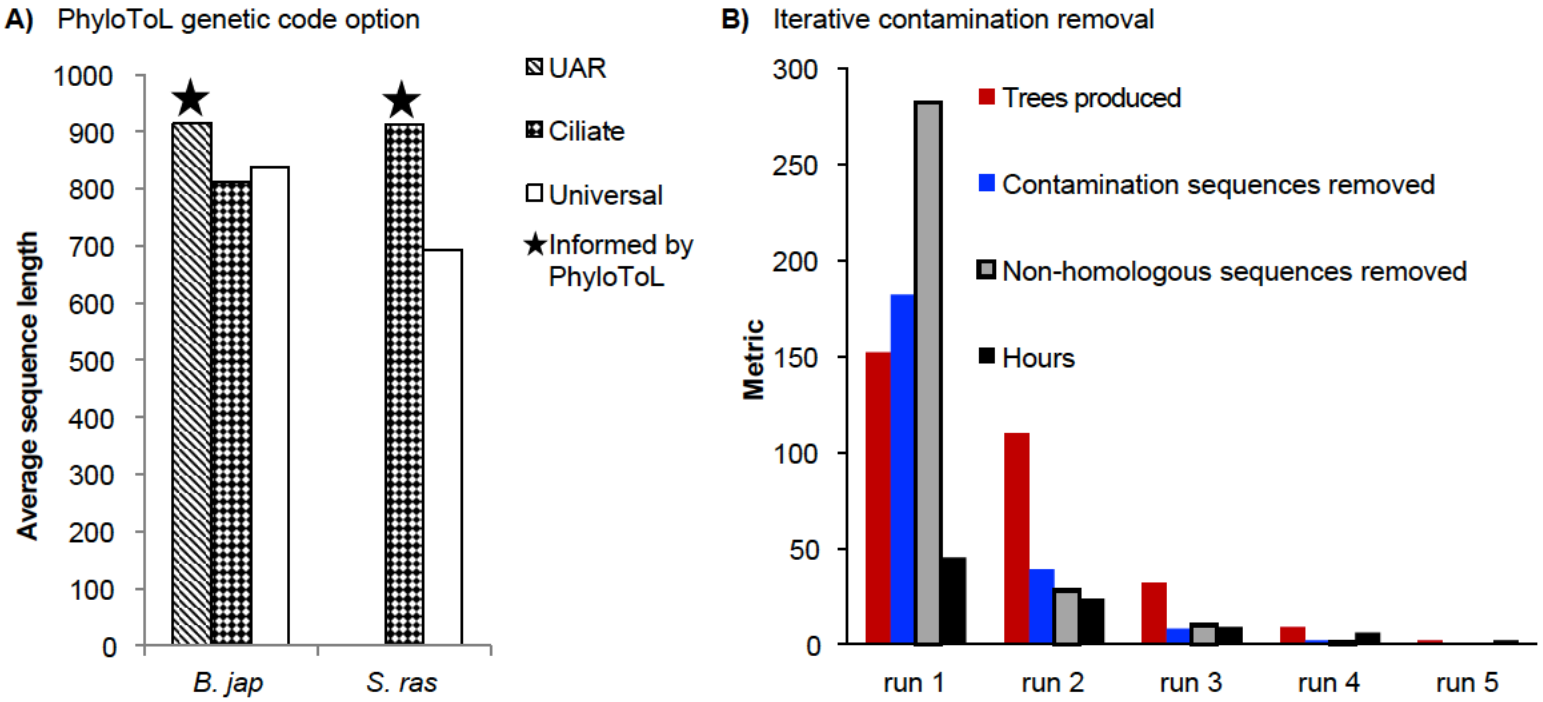

Figure 2.2. Evaluation of performance of the first and second component of PhyloToL (Figures 2.1*A*, 2.1*B*). A) Gene family assessment per taxon performance using the inferred genetic code (indicated with a star) and the ciliate and universal genetic codes for the ciliates *Blepharisma japonicum* and *Strombidium rassoulzadegani*. The length of the inferred sequences is higher when using the informed genetic code because it will not terminate the sequences at potentially reassigned in-frame stop codons. B) Example of contamination removal using our test dataset, containing 152 GFs with up to 167 taxa. Overall it needed 5 iterations to remove all contaminant and non-homologous sequences with most of the sequence removal occurring during the first iteratio

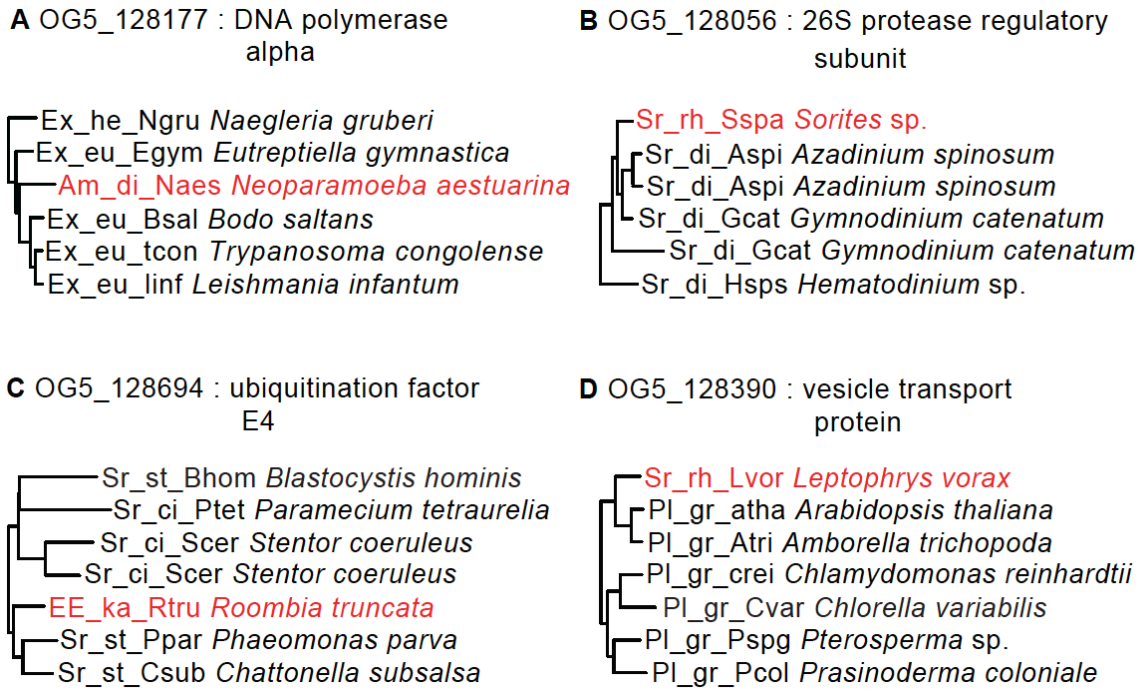

Figure 2.3. Examples of contamination from gene trees, which are used to define rules for the contamination removal loop of component 3 of PhyloToL (See Figure 2.1*C*). All sequences are named by major clade (Am=Amoebozoa, EE = everything else, Ex = Excavata, Pl = Archaeplastida, Sr = SAR), "minor" clade (di = Dinophyceae, he = Heterolobosea, eu = Euglenozoa, st = Stramenopile, ci = Ciliophora, ka = Katablepharidophyta, gr = green algae, rh = Rhizaria) and a four-digit code unique to each species (e.g. Ngru = *Naegleria gruberia*). A) Possible case of contamination in *Neoparamoeba aestuarina* by an endosymbiontic excavate. B) Possible case of contamination in *Sorites* by an endosymbiontic dinoflagellate. C) Possible case of contamination from *Roombia truncata*'s diatom food source. D) Possible case of contamination in *Leptophrys vorax* from its green alga food source.

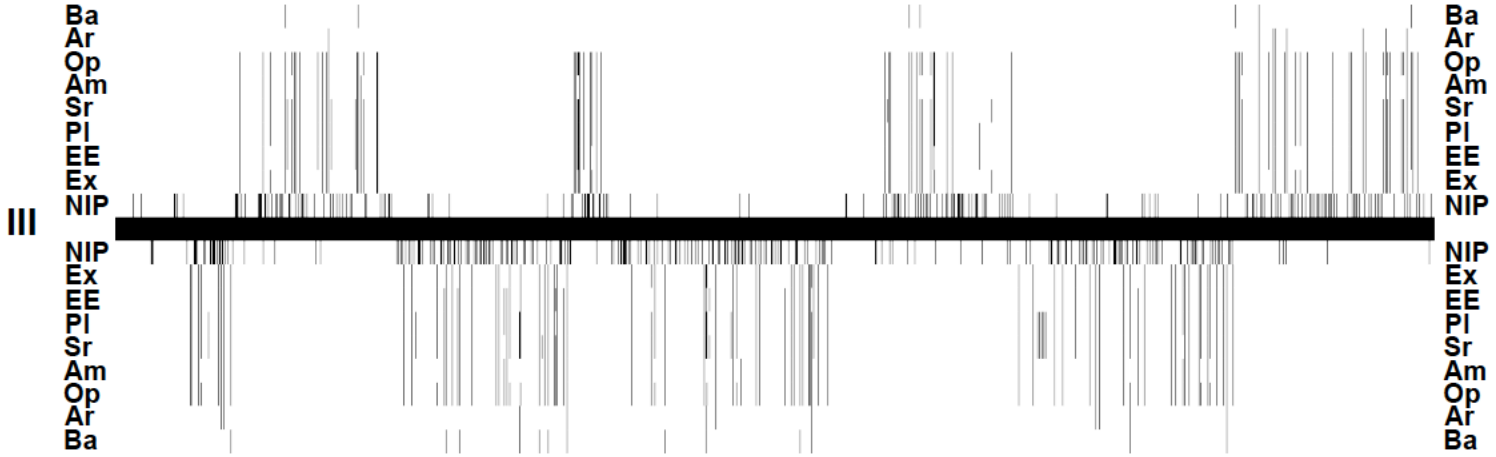

Figure 2.4. Example of phylogenomic map of the chromosome III of *Trypanosoma brucei* generated by combining PhyloToL and PhyloChromoMap (Cerón-Romero, et al. 2018). Horizontal line represent chromosome 3 of *Trypanosoma brucei* and bars above/below reflect levels of conservation. First row from the bottom (NIP, "not in pipeline") indicates ORFs that do not match our criteria for tree inference (i.e. likely *Trypanosoma*-specific, highly divergent and/or misannotated ORFs). The remaining rows (bottom to top) reflect the presence or absence of the gene in the major clades Excavata (Ex), orphans (EE, "everything else"), Archaeplastida (Pl), SAR (Sr), Amoebozoa (Am), Opisthokonta (Op), Archaea (Ar), and Bacteria (Ba). Genes are organized in polycistronic gene clusters (PGC) with variable gene density as described in results/discussion.

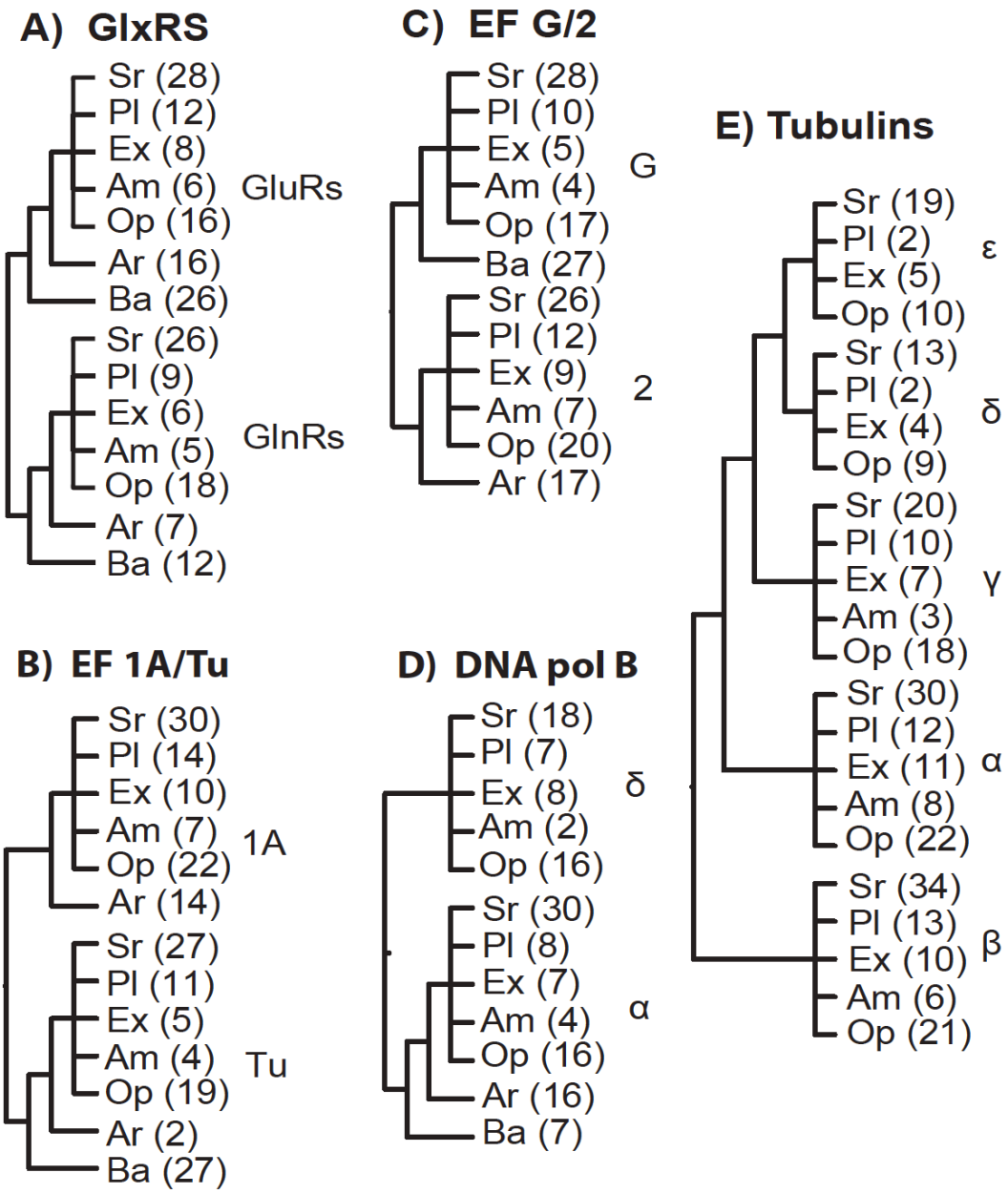

Figure 2.5. PhyloToL homology assessment for well-known GFs that duplicated prior to LUCA. Subfamilies of these ancient GFs are often categorized in different orthologous groups by OrthoMCL. The cartoon trees show the reconstruction of the phylogeny of 5 of the 8 analyzed ancient GF by PhyloToL. A) glutamyl- and glutaminyl-tRNA synthetases, B) elongation factors Tu/1a, C) elongation factors G/2, D) family B DNA polymerase, E) Tubulins. Ar = Archaea, Ba = Bacteria, Op = Opisthokonta, Am = Amoebozoa, Ex = Excavata, Pl = Archaeplastida, Sr = SAR. The number in every tip represents the number of species per major clade.

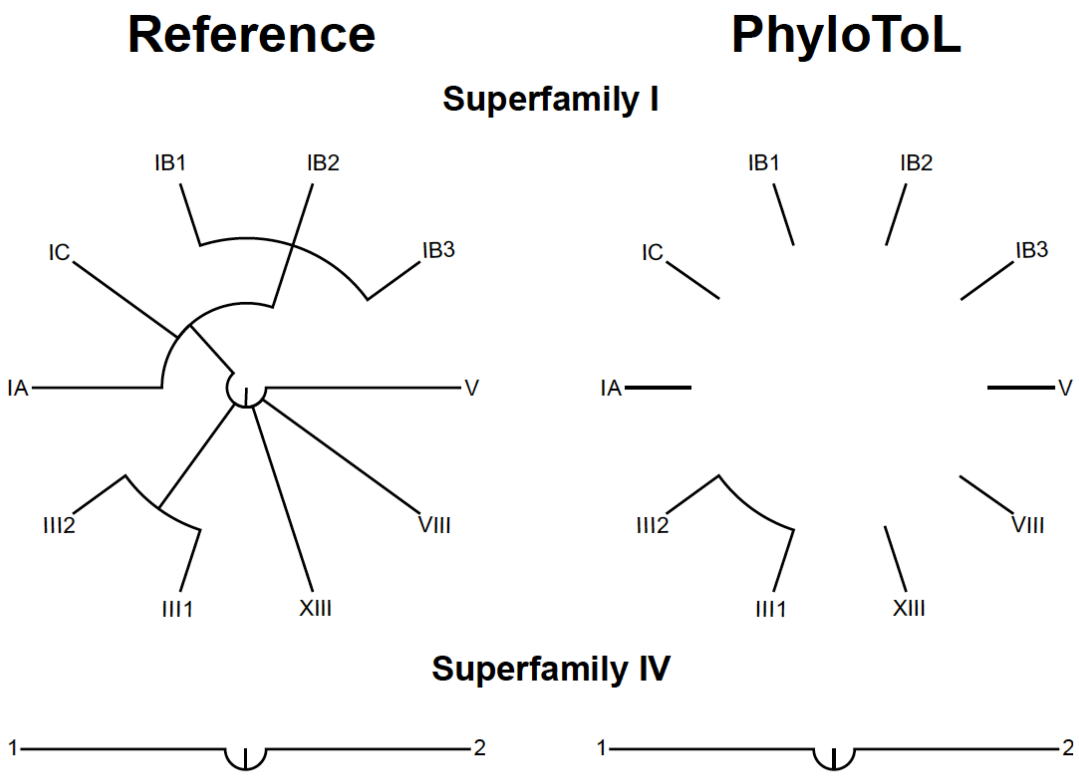

Figure 2.6. PhyloToL homology assessment for candidate superfamilies (S) of outer membrane pore-forming proteins as proposed by Reddy and Saier (2016). The left hand "Reference" columns show the proposed superfamilies SI and SIV while the right hand "PhyloToL" column shows the surviving homologs (i.e. those connected by lines). Only cluster III of SI and the two gene families of SIV are homologous based on PhyloToL's default parameters (i.e. GUIDANCE V2.02: sequences cutoff = 0,3, column cutoff = 0.4, 5 iterations).

CHAPTER 3

PHYLOGENOMIC ANALYSES OF 2,700 GENES IN 150 LINEAGES SUPPORT A ROOT OF THE EUKARYOTIC TREE OF LIFE BETWEEN OPISTHOKONTS (ANIMALS, FUNGI AND THEIR MICROBIAL RELATIVES) AND ALL OTHER LINEAGES[3]

3.1 Abstract

Advances in phylogenetic methods and high throughput sequencing have allowed the reconstruction of deep phylogenetic relationships in the evolutionary history of eukaryotes. Yet, the root of the eukaryotic tree of life remains elusive. The most 'popular' (i.e. in text books and many reviews) hypothesis for the root is between Bikonta (Opisthokonta + Amoebozoa) and Unikonta (all other eukaryotes), which emerged from analyses of a single gene fusion and a limited sampling of eukaryotic lineages. Subsequent highly cited studies based on concatenation of genes supported this hypothesis with some variations or proposed a root between the excavate clade Discoba and all other eukaryotes. Concatenation of genes fails to account for evolutionary events such as gene duplication-loss, incomplete lineage sorting and lateral gene transfer. A more recent study using gene tree-species tree reconciliation methods suggested the root lies between Opisthokonta and all other eukaryotes, but the study included only 59 taxa and 20 genes. Here we apply a gene tree – species tree approach to a gene- and taxon-rich database (i.e. 2,700 gene families from two sets of ~150 diverse eukaryotic lineages) to

[3] Ceron-Romero MA, Fonseca MM, Katz LA. In prep. Phylogenomic analyses of 2,700 genes in 150 lineages support a root of the eukaryotic tree of life between opisthokonts (animals, fungi and their microbial relatives) and all other lineages.

assess the root. Our results estimate a root between Fungi and all other eukaryotes or between Opisthokonta and all other eukaryotes. Finding the root of the eukaryotic tree of life is critical for the field of comparative biology as it allows to understand the timing and mode of evolution of characters across the evolutionary history of eukaryotes.

3.2 Introduction

Among the more controversial topics in the study of the history of life on Earth is the location of the root of the eukaryotic tree of life (EToL), which likely dates to around 1.6-1.8 billion years (de Duve 2007; Parfrey, et al. 2011). While there has been substantial progress on defining major eukaryotic clades such as Archaeplastida, Opisthokonta, SAR and Amoebozoa (Rodriguez-Ezpeleta, et al. 2005; Steenkamp, et al. 2006; Burki, et al. 2007; Hampl, et al. 2009; Adl, et al. 2012; Jackson and Reyes-Prieto 2014; Cavalier-Smith, et al. 2015; Katz and Grant 2015), the location of the root of EToL remains elusive.

Among the more highly-cited hypotheses have been a root within Archezoa (Cavalier-Smith 1989, 1993) or between Unikonta - Bikonta (Stechmann and Cavalier-Smith 2002, 2003; Derelle and Lang 2011; Derelle, et al. 2015). The now-falsified Archezoa root proposed amitochondriate eukaryotes (e.g., microsporidians, diplomonads (e.g. *Giardia*), parabasalids (e.g. *Trichomonas*)) as the earliest-diverging lineages with all other mitochondria-containing lineages radiating after this divergence. This hypothesis lost support when the lack of mitochondria was demonstrated to be a derived character (Roger 1999).

In the past two decades, the Unikonta - Bikonta root has gained popularity and can be found in many text books. Though both clades have incorporated numerous taxonomic changes over the years, the root was first articulated as being between Opisthokonta + Amoebozoa and the rest of the eukaryotes (Stechmann and Cavalier-Smith 2003). More recently, a new clade including Unikonta and former bikont lineages (i.e. Apusozoa, Breviata) was defined as Amorphea (Adl, et al. 2012) with the root dividing Amorphea and the remaining eukaryotes (Derelle, et al. 2015).

Advances in high-throughput sequencing technologies allow better estimation of eukaryotic phylogeny by providing the opportunity to explore bigger datasets and include non-model organisms such as the rhizarians *Quinqueloculina* or the glaucophyte *Gloeochaete* (Burki, et al. 2007; Jackson and Reyes-Prieto 2014; Katz and Grant 2015; Brown, et al. 2018). A popular approach to take advantage of such opportunities is by inferring phylogenies from supermatrices by concatenating multiple genes in a single alignment (Rodriguez-Ezpeleta, et al. 2005; Dunn, et al. 2008; Wickett, et al. 2014; Derelle, et al. 2015). Analyses of multiple concatenated eukaryotic genes of putatively bacterial origin (i.e. mitochondrial) have either supported the Unikonta-Bikonta root (Derelle and Lang 2011; Derelle, et al. 2015) or suggested a new root between Discoba (Excavata) and the other eukaryotes (He, et al. 2014).

Alternative methods have supported diverse root possibilities. For instance, a genome-wide analysis of rare genomic changes suggests a root between Archaeplastida and the other eukaryotes (Rogozin, et al. 2009), and an analysis based on the presence/absence of an encounter structure for the endoplasmic reticulum and the mitochondria suggests a root between Amorphea + Excavata and the rest of eukaryotes

(Wideman, et al. 2013). A promising method for species tree inference is gene tree parsimony (GTP), which not only takes advantage of the power of gene-rich databases but also considers gene duplications and losses across individual gene trees. Based on only 20 gene trees, a preliminary GTP analysis estimated a root between Opisthokonta and the rest of eukaryotes (Katz, et al. 2012), which is consistent with initial analysis of the fusion between dihydrofolate reductase (DHFR) and thymidylate synthase (TS) genes (Stechmann and Cavalier-Smith 2002).

Phylogenomic methods vary in their approach to identify and account for evolutionary events such as lateral gene transfer (LGT), gene transfer from endosymbiosis (EGT) and gene duplications/losses, which can be prevalent in many eukaryotic lineages (Galtier and Daubin 2008; Burki, et al. 2014; Katz 2015; Panchy, et al. 2016). Supermatrix methods require identifying and removing paralog sequences before building the concatenated alignment. Yet, distinguishing orthology from paralogy can be very difficult, particularly at scales of >1 billion years of eukaryotic evolution. Despite the limitations of supermatrix methods, which discard informative data (e.g. gene duplications and losses), their tractability has made them popular choices in studies estimating the root of EToL.

There are also alternative methods that estimate the best species tree by minimizing the discordance between candidate species trees and a set of gene trees. In contrast to supermatrix methods, these gene tree – species tree reconciliation methods allow the incorporation of informative data from different evolutionary events. Some of these methods assume that the discordance between gene trees and species tree is due to either incomplete lineage sorting (Mirarab, et al. 2014; Mirarab and Warnow 2015), gene

duplication and loss (Chaudhary, et al. 2010) or LGT (Whidden, et al. 2014). Other reconciliation methods consider multiple evolutionary events at once (De Oliveira Martins, et al. 2016; Mallo and Posada 2016), which substantially increases the needs for computational power.

Here we apply an approach based on the reconciliation of gene and species trees to infer the root of EToL and evaluate the levels of support for the different published hypotheses. For this purpose, we use the recently published phylogenomic pipeline PhyloToL (Ceron-Romero, et al. 2019) and build a database of phylogenetic trees from 2786 gene families including 150 species distributed across the whole EToL.

3.3 Results

3.3.1 Building the phylogenomic datasets

Using our taxon- and gene-rich phylogenomic pipeline, PhyloToL (Ceron-Romero, et al. 2019), we built two datasets that each include 2,786 gene families and ~160 species from 140 and 158 genera (Table 3.1). The two datasets varied based on taxon selection criteria: for the 'SEL+' dataset, we selected representative species within clades based on our assessment of data quality and taxonomic breadth; and for the 'RAN+' dataset, we randomly chose even numbers of species among the major eukaryotic clades (i.e. Opisthokonta, Amoebozoa, Archaeplastida, Excavata, SAR and some orphan lineages (Table 3.1). We also generated two additional databases by excluding the fast-evolving Microsporidia (i.e. SEL- and RAN-) as inclusion of these lineages can generate phylogenetic artifacts such as long-branch attraction (Embley and Hirt 1998; Hirt, et al. 1999; Van de Peer, et al. 2000). We chose 2,786 gene families from

among ~13000 gene families in PhyLoToL, selecting genes that before iterative alignments are present in at least 25 taxa of at least 4 major eukaryotic clades (see methods).

3.3.2 Inference on location of the root

Though we set out to deploy two gene tree – species tree reconciliation methods to infer the root of the eukaryotic tree of life, we were constrained to focus on only one for the analyses presented here. Our original intent was to use both a Bayesian supertree approach with the software *guenomu* (de Oliveira Martins et al., 2016) and a gene tree parsimony approach with the software package iGTP (Chaudhary et al. 2010). Both approaches are appropriate when species have multiple copies of any given gene as both account for duplications and losses. *Guenomu* addresses the disagreement between gene trees and the species tree in a jointly/multivariate manner, assuming that the source of disagreement is a composition of duplication and losses, incomplete lineage sorting, LGT, or other stochastic processes (De Oliveira Martins, et al. 2016). On the other hand, iGTP assumes that the disagreement between gene and species tree is only due to either duplication, duplication-loss, or deep coalescence. Unfortunately, *guenomu* failed to converge in an estimate of species trees after being run for multiple weeks on an HPC, likely due to the complexity of the data, so we continued only with iGTP.

Using iGTP, we estimated the most parsimonious rooted tree of eukaryotes for each of our four datasets, all of which indicate Fungi as the earliest branching group (Figure 3.1). Other less parsimonious but frequent alternatives indicate glaucophytes or the apusozoan *Fabomonas tropica* as the earliest branching group or taxon. Across all

repetitions of the analysis, the most frequent following branching group is the opisthokonts (i.e. the other opisthokonts when the earliest branching group was Fungi). These results leave open the possibility of a root between Opisthokonta and the other eukaryotes but with some factor such as LGT or missing data influencing iGTP calculations.

3.3.3 Comparison to published hypotheses

We also used iGTP to evaluate various hypotheses from the literature including a root: between Opisthokonta and others (Stechmann and Cavalier-Smith 2002; Katz, et al. 2012), between Discoba (Excavata) and others (He, et al. 2014), the Unikonta – Bikonta root (Stechmann and Cavalier-Smith 2003; Derelle, et al. 2015), and an alternative root (Ancyromonadida + Metamonada) – others. Here we estimate the reconciliation cost of a species tree given a constrained topology to reflect the different hypotheses of the root of EToL (Figure 3.2, x-axis). In addition to these 4 hypotheses, we also calculated and compared the reconciliation cost of a species tree reflecting our initial estimates, placing the root between Fungi and the other eukaryotes. The results show that for the datasets SEL- and RAN- our inferred root of Fungi + others is more parsimonious than the other 4 hypotheses, while for dataset SEL+ and RAN+ the most parsimonious root is Opisthokonta + others (Figure 3.2).

To assess the difference in reconciliation, we conducted pairwise t-tests among all 4 hypotheses in all datasets. Our results show that for datasets SEL+, SEL- and RAN+ there are not significant differences between Opisthokonta + others and Fungi + others (t-student, $p > 0.01$, Table S6), while the root between Fungi and the rest of eukaryotes was

significantly more parsimonious than the remaining hypothetical roots (t-student, $p < 0.01$; Figures 3.2, S6). For the dataset RAN- the root Fungi + others was more significantly parsimonious than all other hypotheses.

3.4 Discussion

This study analyzes 2,786 gene trees with four taxon samplings including ~200 diverse eukaryotic taxa, perhaps the largest analysis yet to address the root of the eukaryotic tree of life. As in Katz, et al. (2012), we used gene tree parsimony as implemented in the software iGTP to estimate the root of EToL that minimizes gene duplications and loses. Given the importance of gene duplication/loss for the evolution of eukaryotic genomes (e.g. Wolfe 1997; Otto and Whitton 2000; Dehal and Boore 2005), their inclusion in the estimation of the most likely root of EToL represents a powerful alternative to studies that are based on a supermatrix approach (Guigo, et al. 1996; Chaudhary, et al. 2010), as the latter require users to discard potentially-informative paralogs.

Across our analyses we find that the root with the best reconciliation cost is either with the Fungi or Opisthokonta as sister taxon to all other eukaryotes. The Fungi + others root is consistently the most supported root regardless of which dataset is used in the analysis (Figure 3.1). This hypothesis was previously discussed based on the fact that Fungi have osmotrophic feeding while all other eukaryotes have phagotrophic feeding (Martin, et al. 2003). Moreover, fungi contain more ATP pathways than any other major eukaryotic clade, including for ATP synthesis under anoxic and high sulfide conditions that resemble the environment on early eukaryotic evolution. Advances in the analysis of

fossil record are also very promising. For instance, a new fossil was found in Arctic Canada, which is as twice as old as the fossil used for the current estimates of the origin of fungi (Loron, et al. 2019). Many other pre-Ediacaran fossils also look more similar to fungi than to any other clade but much more work needs to be done to classify them as Fungi (Butterfield 2005, 2009). Although there are not previous phylogenetic studies to support that Fungi is the earliest branching eukaryotic clade and the monophyly of Opisthokonta is widely accepted (Baldauf and Palmer 1993), these fossil record findings and the characteristics of energy production in fungi encourage further exploration of this hypothesis.

Our comparison of hypotheses shows that Opisthokonta + others has similar support as Fungi + others. Opisthokonta + others was demonstrated in previous studies also using gene tree parsimony (Katz, et al. 2012) and was originally proposed based on DHFR-TS fusion gene (Stechmann and Cavalier-Smith 2002), though the gene fusion had a more complex distribution upon additional taxon sampling. Our results open up the possibility that Opisthokonta + others is the actual root, while Fungi + others is a phylogenetic artifact due to either LGT or high rates of gene loss. We found only an insignificant number of potential LGT event between Bacteria and Fungi in our databases. However, our data is comprised of protein sequences, which makes it difficult to track LGT in highly conserved genes across the tree of life, and there is always the possibility that PhyloToL's database is lacking some key bacteria to uncover those LGT events. Also, Fungi have experienced substantially higher rates of gene loss than other Opisthokonta, which is reflected in their much-reduced genome sizes (Figure S10). If Opisthokonta + others is the actual root, genes that are conserved between Opisthokonta

and the other eukaryotes but independently lost in Fungi, could be considered by iGTP as phylogenetic information to put Fungi in many cases at the root while putting the other Opisthokonta closer to the other eukaryotes. This outcome would be even more likely if Opisthokonta experienced frequent genome duplication events and many of the genes kept in Fungi came from different paralogs than the ones kept in the other Opisthokonta.

In a limited number of analyses, we found a surprising root of Glaucophytes + others (Glaucophytes, (Opisthokonta, others)), which appears consistently as one the most parsimonious roots (always less parsimonious than Fungi + others) in all datasets (Figure 3.1). Given that Glaucophytes are the minor more poorly represented in the gene trees (Figure S11), this seems to be the same potential artifact caused by high rates of gene loss that we described for Fungi. However, in this case, the lack of genes is due to incomplete sequencing instead of high rates of gene loss, but the outcome is the same: a whole clade with substantially fewer genes than their closest relatives (i.e. the other Archaeplastida). Previous studies have shown that the gene tree parsimony approach for species tree inference is sensitive to missing data (Burleigh, et al. 2011; Davis, et al. 2019). Given that here we are using a duplication/loss model it is likely that missing data, particularly when all involved taxa from the same clade, influenced the inferences by undermining calculations of gene losses.

An important issue in analyses of the root of EToL has been the inconsistency in the definition of taxa in studies based on the supermatrix approach (Derelle and Lang 2011; He, et al. 2014; Derelle, et al. 2015; Brown, et al. 2018). Most of these studies support a Unikonta-Bikonta root but propose taxonomic changes for the Unikonta group. Even when He, et al. (2014), also using a supermatrix approach, proposed a root in

Excavata, this root was later re-analyzed concluding that the data supports the Unikonta-Bikonta root (Derelle, et al. 2015). The lack of consistency that results from taxa and gene sampling could be explained by the limitations of the supermatrix approach. For instance, choosing orthologs in "orphan" lineages such as ancyromonads could be a huge source of bias or noise. Also striking is the fact that all other studies that use alternative methods to supermatrix always predict a different root than the Unikonta-Bikonta (Stechmann and Cavalier-Smith 2002; Rogozin, et al. 2009; Katz, et al. 2012; Wideman, et al. 2013).

There are many caveats when exploring the root EToL. It is expected that LGT, incomplete lineage sorting as well as duplications and losses play a role in the phylogenetic history of eukaryotic genomes. While ideally all these evolutionary factors would be considered in phylogenomic studies, their incorporation increases significantly the complexity of the analyses and the computation needs. Currently, the only gene tree – species tree reconciliation tool demonstrated to consider all these evolutionary factors for species tree inference is *guenomu*. However, this tool does not support the ~1.8 billion years of evolution represented in our databases. In order to deal with the complexity in our databases, we decided to focus only on duplications and losses. Given the deep divergences represented in our databases, incomplete lineage sorting is expected to have a small impact. Most LGT events in eukaryotic genomes come from organelles of prokaryotic origin. There is evidence that ancient interdomain LGT events are rare, with the exception of those coming from plastids (Katz 2015). Given the lack of gene tree – species tree reconciliation tools for species tree inference that support the level of divergence in our data and that considers a combined effect between LGT and

duplications/losses, we decided to filter possible LGT events before and during alignment building. Despite these caveats, the diversity represented in this study, the more phylogenetically informative approach based on gene tree parsimony, and the consistent results despite changes in taxa selection, show the robustness of our analyses and results.

3.5 Methods

3.5.1 Taxa selection

We started with the database of PhyloToL, which contains 1007 taxa including Bacteria, Archaea and Eukaryotes. From this database, we generated two subsets of 155 eukaryotic taxa with two different criteria: 1) selecting taxa based on maximizing the inclusion of eukaryotic clades and the quality of the data (SEL+) and 2) selecting taxa randomly among the major eukaryotic clades Opisthokonta, Amoebozoa, Archaeplastida, Excavata, SAR and some orphan lineages (RAN+; Table 3.1). We also generated two extra datasets without microsporidians (SEL- and RAN-) in order to account on a possible effect over the phylogenetic inferences due to microsporidians fast-evolutionary rates.

3.5.2 Gene family selection

PhyloToL contains 13104 protein-coding gene families. We chose the gene families that contain at least 25 taxa representing at least 4 of the 5 major eukaryotic clades. Additionally, at least 2 of the major clades had to contain at least 2 minor clades (e.g. Glaucophytes and Rhodophyta are minor clades in the major clade Archaeplastida). We produced an alignment and a phylogenetic tree for each gene family and filtered the gene families that are exclusive of eukaryotes or the ones in which eukaryotes were

monophyletic. From a total of 3002 gene families that met our criteria, 2786 passed the initial steps of PhyloToL when including only the data from the dataset SEL+. This 2786 GFs were used for further analyses with all datasets.

3.5.3 Root inference

In order to infer the root of the EToL, we use two supertree tools for species tree inference, the Bayesian-based *guenomu* and the gene tree parsimony tool iGTP. While iGTP considers that the discrepancy between gene trees and species tree is due to either duplications, duplications-losses or deep coalescence; *guenomu* considers jointly the effect of all these and other evolutionary processes. We ran guenomu with gene trees produced with MrBayes (Huelsenbeck and Ronquist 2001; Ronquist, et al. 2012) using the dataset SEL+. MrBayes was run with four Markov chains incrementally heated with the default values and each chain started with a randomly generated tree and ran for $1x10^7$ generations. For every 100 generations, one tree was sampled for the analysis. The posterior distribution of trees, after discarding the first 25% as burn-in, was summarized in a 50% majority-rule consensus tree. Two independent replicates were conducted and inspected for consistency. We did not get a solution in a reasonable time; therefore, we chose not to continue with *guenomu* and continued further analyses just with iGTP.

We ran iGTP for the four datasets with gene trees produced with RAxML v.8.2.4 (Stamatakis 2014) with 10 ML searches for best-ML tree (option "-# 10") using rapid hill-climbing algorithm (option "-f d") and no bootstrap replicates. The protein evolution model used was evaluated during the gene tree inference (option "-m PROTCATAUTO") by testing all models available in RAxML (e.g. JTT, LG, WAG, etc) with optimization of

substitution rates and of site-specific evolutionary rates which were categorized into four distinct rate categories for greater computational efficiency. In the implementation of iGTP, we decided to increase the accuracy by running 100 replicates per dataset. However, in preliminary analyses we detected that the root of the input gene trees and their order in the 100 replicates could impact the results in iGTP, therefore we randomly re-rooted gene trees and randomly shuffled the order of the trees in each replicate.

3.5.4 Comparing different root hypotheses

For the datasets SEL+, RAN+, SEL- and RAN-, we compare 5 different hypotheses of the root of EToL. These hypotheses are: 1) the most parsimonious root according to the previous analysis, 2) between Opisthokonta and the rest of eukaryotes, 3) between Discoba (Excavata) and rest of the eukayotes, 4) between Unikonta and Bikonta, and 5) between Metamonada (Excavata) + Ancyromonadida and the rest of eukaryotes. For the Unikonta-Bikonta root, different alternatives were evaluated according to the multiple changes on the definition of the Unikonta clades, but only the best alternative was used for further comparisons. In order to compare the hypotheses, we constrained species trees according to every hypothesis and calculated the reconciliation cost per hypotheses in each dataset.

Table 3.1. A summary of taxon selection for each dataset. Genera in **bold** are only in the taxonomy informed selected datasets (i.e. SEL+), underlined genera are only in the randomly selected within clades datasets (i.e. RAN+). The genera in red are microsporidians, which we excluded from datasets SEL- and RAN- because they often fall on very long branches (Embley and Hirt 1998; Hirt, et al. 1999; Van de Peer, et al. 2000). The numbers represent the amount of species included and the number of whole genomes in parenthesis.

		Taxa (genomes)	
Major Clade	**Genera**	**SEL+**	*RAN+*
Amoebozoa	*Acanthamoeba, Acytostelium, **Clydonella**, Dictyostelium, Endostelium, Entamoeba, Filamoeba, Flamella, Gocevia, **Hartmanella,** Mastigamoeba, Mayorella, **Neoparamoeba**, **Ovalopodium**, Paramoeba, **Parvamoeba**, Pessonella, **Physarum**, Polysphondylium, **Stenamoeba**, Stereomyxa, Thecamoeba, Unda,Vannella, Vermistella, Vexillifera*	22(3)	23(4)
Fungi	*Aspergillus, **Batrachochytrium**, Candida, Cryptococcus, Dacryopinax, Encephalitozoon*, Enterocytozoon*, Laccaria, Malassezia, Melampsora, **Nematocida***, **Neurospora**, **Nosema***, Phanerochaete, Piromyces, Puccinia, **Rhizophagus**, **Saccharomyces**, **Schizosaccharomyces***	13(11)	13(10)
Other Opisthokonta	***Amphimedon**, Anopheles, Apis, **Aplysia**, Branchiostoma, **Caenorhabditis**, **Capitella**, Capsaspora, Carteriospongia, **Ciona**, Culex, **Drosophila**, Equus, **Fonticula**, Gallus, Helobdella, **Homo**, Hydra, Hydractinia, Leucetta, Lubomirskia, Macaca, **Mnemiopsis**, Monosiga, **Nematostella**, Oikopleura, Ornithorhynchus, Oscarella, Pan, **Pleurobrachia**, Rattus, **Saccoglossus**, **Salpingoeca**, **Schistosoma**, **Sphaeroforma**, Trichinella, **Trichoplax***	21(12)	21(14)
Archaeplastida	***Amborella**, Arabidopsis, Bathycoccus, Chlorella, **Chondrus**, **Coleochaete**, **Compsopogon**, Crustomastix, Cyanidioschyzon, **Cyanophora**, Cyanoptyche, Erythrolobus, **Galdieria**, Glaucocystis, Mantoniella, Mesostigma, Micromonas, **Nephroselmis**, Ostreococcus, **Physcomitrella**, Picochlorum, Picocystis, **Porphyra**, Porphyridium, Pycnococcus, **Rhodella**, Rhodosorus, **Ricinus**, **Volvox***	20(7)	18(4)
SAR	*Alexandrium, **Ammonia**, Amphidinium, Amphiprora, Amphora, Astrosyne, **Aureococcus**, **Bigelowiella**, **Blastocystis**, **Bolidomonas**, Brandtodinium, Brevimastigomonas, Bulimina, **Cafeteria**, **Chattonella**, Chlorarachnion, **Chrysoreinhardia**, Corallomyxa, Corethron, Cryptosporidium, Ectocarpus, Eimeria, **Euglypha**, Euplotes, **Extubocellulus**, Florenciella, Fragilariopsis, Fucus, Gonyaulax, Gregarina, **Gymnodinium**, Gymnophrys, Karlodinium, Lankesteria, **Leptophrys**, Lingulodinium, Lotharella, Nannochloropsis, Nitzschia, Ochromonas, Oxytricha, Paracercomonas, Pelagodinium, **Perkinsus,** **Phaeodactylum**, **Phaeomonas**, Phyllostaurus, **Phytophthora**, **Plasmodium**, **Pyrodinium**, Pythium, **Reticulomyxa**, Rhizochromulina, **Saprolegnia**, Sarcinochrysis, Scrippsiella, **Sorites**, **Spumella**, Stylonychia, Synchroma, Tetrahymena, **Thalassionema**, **Thalassiosira**, **Thraustochytrium**, **Toxoplasma**, **Vitrella***	40(17)	39(7)
Excavata	***Euglena**, **Eutreptiella**, Giardia, Histiona, Histomonas, Jakoba, Leishmania, **Malawimonas**, **Monocercomonoides**, **Naegleria**, Neobodo, **Percolomonas**, **Reclinomonas**, Sawyeria, Seculamonas, Spironucleus, Stachyamoeba, Strigomonas, **Trichomonas,** Trimastix, **Tritrichomonas**, Trypanosoma*	22(7)	21(12)
Other eukaryotes	*Acanthocystis, Calcidiscus, **Choanocystis**, **Chrysochromulina**, Chrysoculter, **Collodictyon**, Cryptomonas, **Diphylleia,** Emiliania, Fabomonas, Goniomonas, Hanusia, Hemiselmis, Isochrysis, **Palpitomonas**, Pavlova, Phaeocystis, **Pleurochrysis**, Prymnesium, Raphidiophrys, Rhodomonas, Rigifila, **Roombia**, **Subulatomonas,** Telonema, Thecamonas, Tsukubamonas*	20(1)	20(1)

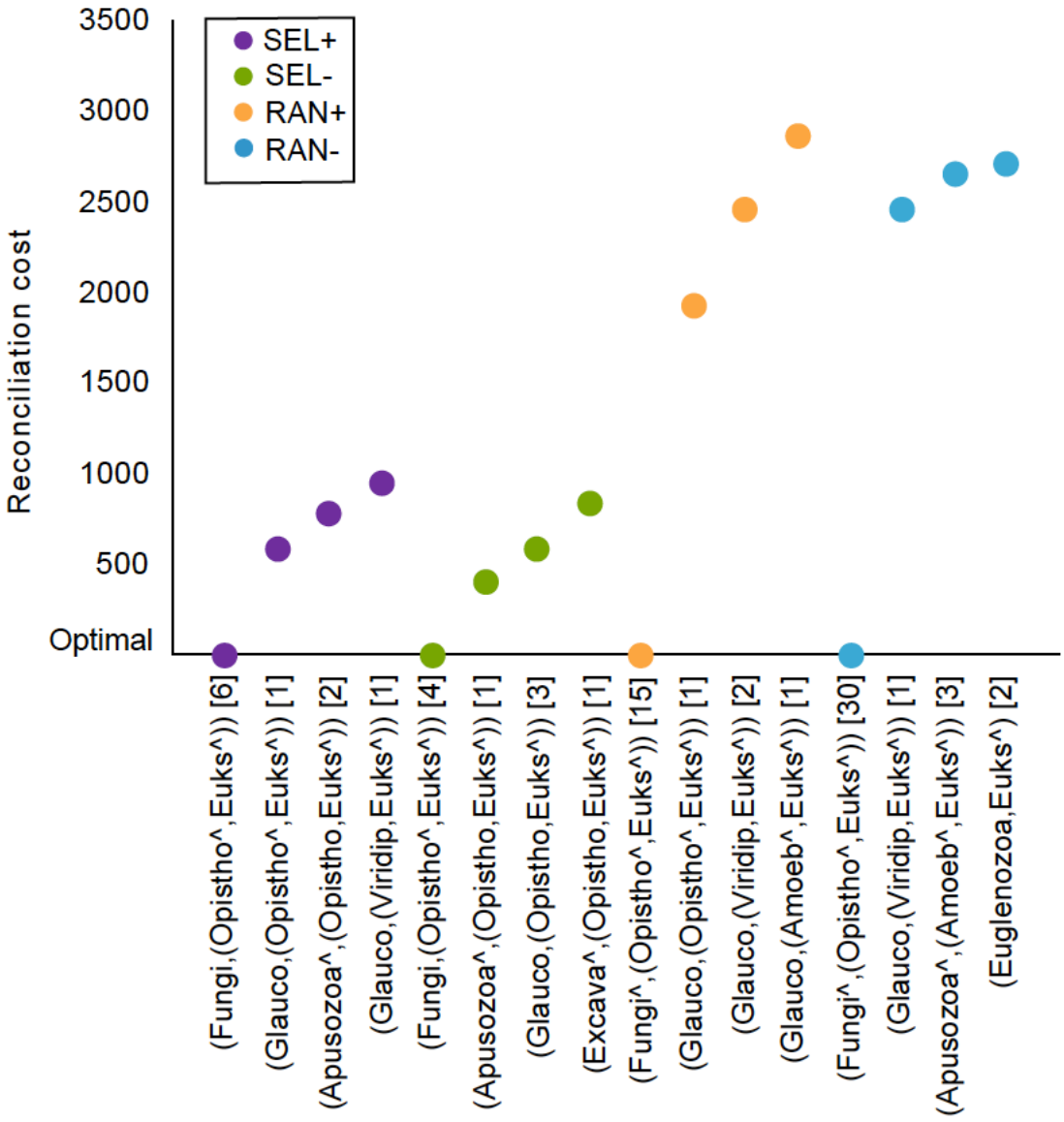

Figure 3.1. A root between fungi and all other eukaryotes is the most parsimonious hypothesis inferred from 100 iterations for each of our four datasets. SEL+: selected taxa including microsporidians, SEL-: selected taxa excluding microsporidians, RAN+: random within major clades and including microsporidians, RAN-: random within major clades and excluding microsporidians (More details are in Table 3.1). Here we report the four most parsimonious topologies (reconciliation cost is relative to the optimal/lowest value) in the 100 iterations. Each of the four most parsimonious topologies could appear multiple times in the 100 iterations. The number in brackets is the consecutive times that the topology first appears in a ranking of reconciliation cost values out of the 100 iterations. The caret (^) implies no monophyly. In datasets SEL+ and RAN+ the microsporidians do not fall in the same clade as the rest of opisthokonts. In RAN+ and RAN- the best species trees have Fungi as not-monophyletic as separating *Piromyces* from the other Fungi.

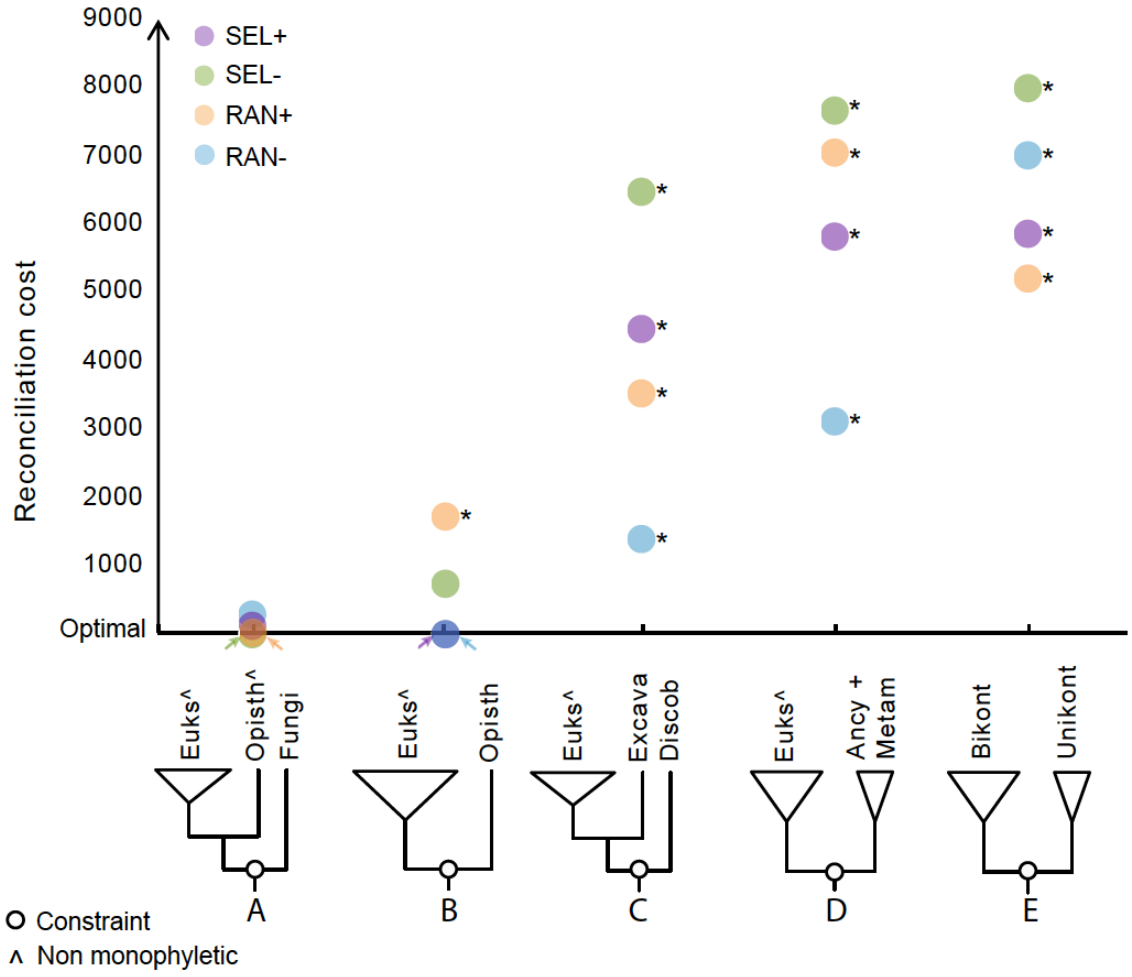

Figure 3.2. Comparison of five hypotheses for the root from the literature estimated using iGTP with the 4 datasets (repetitions). We constrained the species trees according to each hypothesis and estimate the reconciliation costs, showing the costs relative to the optimum for each dataset (the lowest value). The five hypotheses here are: A) between fungi and the others (our estimate from the previous analysis), B) between Opisthokonta and the others (Stechmann and Cavalier-Smith 2002; Katz, et al. 2012), C) between Ancyromonadida + Metamonada and the others, D) between Discoba and the others (He, et al. 2014), and E) between unikonta and bikonta (Stechmann and Cavalier-Smith 2002; Derelle, et al. 2015). The empty circle indicates where in the tree the constrain was applied and other notations are as in Figure 3.1. The reconciliation cost of fungi + others is significantly different to the reconciliation costs in all other hypotheses except Opisthokonta + others in SEL+, SEL- and RAN- (t-student, $p > 0.001$; more details about statistical tests in Table S6).

APPENDIX A

SUPPLEMENTARY TABLES

Table S1. Size of young region in chromosomes of *P. falciparum*[a]

Size	Chr	Start	End	Chromosome region
218000	4	1	218000	Subtelomeric
191000	1	453001	644000	Subtelomeric
190000	7	1	190000	Subtelomeric
173000	2	775001	948000	Subtelomeric
168000	9	1374001	1542000	Subtelomeric
165000	12	2107001	2272000	Subtelomeric
162000	8	1258001	1420000	Subtelomeric
160000	7	1342001	1502000	Subtelomeric
151000	4	1054001	1205000	Subtelomeric
145000	13	1	145000	Subtelomeric
141000	7	561001	702000	Internal
135000	2	1	135000	Subtelomeric
133000	9	1	133000	Subtelomeric
132000	1	1	132000	Subtelomeric
132000	14	3160001	3292000	Subtelomeric
127000	6	1292001	1419000	Subtelomeric
125000	10	1563001	1688000	Subtelomeric
121000	10	1	121000	Subtelomeric
121000	13	2775001	2896000	Subtelomeric
120000	8	1	120000	Subtelomeric
119000	4	219001	338000	Internal
114000	8	404001	518000	Internal
114000	11	1	114000	Subtelomeric
108000	4	918001	1026000	Internal
104000	3	1	104000	Subtelomeric
103000	11	1898001	2001000	Subtelomeric
101000	6	1	101000	Subtelomeric
98000	12	1	98000	Subtelomeric
94000	13	1371001	1465000	Internal
92000	14	1	92000	Subtelomeric
91000	5	1	91000	Subtelomeric
91000	12	1683001	1774000	Internal
91000	13	1093001	1184000	Internal
91000	13	2049001	2140000	Internal
86000	5	1258001	1344000	Subtelomeric
85000	3	976001	1061000	Subtelomeric

[a] We define young regions as containing genes in two or fewer major eukaryotic clades, allowing for a single 'interrupting' gene. We only considered internal young regions larger than 90 kb.

Table S2. Characteristics of putative centromeres in chromosomes of *P. falciparum*.

Chr	Size (Kbp)[a]	AT (%)	Gap between genes (Kbp)[b]	Nearest genes
1	2	98	6	*PFA_0585w* and *PFA_0590w*
2[c]	2	97	7	*PFB0490c* and *PFB0495w*
3[c]	2	97	13	*PFC0610c* and *PFC0615w*
4	2	97	6	*PFD0690c* and *PFD0692c*
5	3	94	4	*MAL5_tRNA_Leu1* and *PFC0615w*
6	2	98	7	*PFF0560c* and *PFF0565c*
7	2	98	5	*PfEST* and *PfCRMP2*
8	3	94	3	*PF08_0118* and *MAL8P1.200*
9	2	96	5	*PFI1500w* and *PFI1835c*
10	1	94	3	*PF10_0114* and *PF10_0115*
11	2	98	3	*PF11_0226* and *PF11_0227*
12	2	98	4	*PFL1505* and *PFL1510c*
13	3	94	6	*PF13_0157* and *MAL13P1.151*
14	2	98	5	*PF14_0252* and *PF14_0253*

[a] The sizes are approximations based on the AT content
[b] Gap between genes in which the centromere is residing. The number represents the distance between the 2 nearest genes.
[c] Previously described (Bowman, et al. 1999; Hall, et al. 2002)

Table S3. Genes of *P. falciparum* that were likely transferred through interdomain LGT.

Type	Chr	Accession	Protein	Important characteristics
O	7	XP_002808799	1-cys peroxiredoxin	Apicoplast, response to oxidative stress
O	8	XP_002808807	Ubiquitin-like protease 1	Post-translational
O	8	XP_002808852	GTPase	Vesicles transport, signal transduction, cell cycle control
O	9	XP_001352190	Peptide release factor*	Termination of translation
O	9	XP_001351950	Apicoplast ribosomal protein L35 precursor	Apicoplast, translation
O	4	XP_001351509	Holo-(acyl-carrier protein) synthase*	Activation of ACP for fatty acid synthesis in apicoplast
O	2	XP_001349551	5'-3' exonuclease, N-terminal resolvase-like domain*	Non globular domain inserted in globular domain
O	9	XP_001352042	N-glycosylase/DNA lyase*	Likely involved in DNA repair
O	3	XP_001351267	ABC transporter*	Likely involved in drug resistance
R	5	XP_001351573	Interspersed repeat antigen*	Drug resistance

(O) Old, (R) Recent, (*) Putative

Table S4. Comparison of features among PhyloToL, OneTwoTree (Drori, et al. 2018), SUPERSMART (Antonelli, et al. 2017) and PhyloTA (Sanderson, et al. 2008).

Feature	PhyloToL	OneTwoTree	SUPERSMART	PhyloTA
Scope of study	GF or phylogeny for any species using molecular data from databases or user input	GF or phylogeny of well annotated species using data from GenBank	Incorporate fossil and population genetic data into phylogeny of closely related taxa (shallow nodes)	GF of well annotated species from GenBank
Special features	Highly modular and flexible	Flexible outgroup selection	Advanced dating options	Easy integration with other databases
Data type	Focused on amino acids inferred from DNA	Focused on DNA	DNA and fossil record	Focused on DNA
Markers / GFs	Defined by user according to seed GF database (default = orthoMCL)	Built *de-novo* from GenBank data	Predefined by PhyloTA	Built *de-novo* from GenBank data
Homology calling	Iterative multisequence comparison using GUIDANCE after mapping to OrthoMCL	Markov clustering using OrthoMCL-based algorithm	Initial clustering based on taxonomy, then pairwise sequence comparison	Single-linkage clustering using BLINK
Contamination detection & removal	yes	no	no	no
Orthology calling	Based on gene tree topology or easy export for 3rd party tool	Based on sequence comparison (OrthoMCL-based algorithm)	no	Based on gene tree topology and K-H statistical test
Phylogeny inference	ML or easy export for 3rd party tool	ML or Bayesian, dated	ML or Bayesian, dated	Parsimony
Products	GFs MSAs Gene trees supermatrix sps trees	GFs MSAs Gene trees supermatrix sps trees	GFs MSAs Gene trees sps trees	GFs MSAs Gene trees

Table S5. PhyloToL homology test for candidate superfamilies proposed by Reddy and Saier (2016).

Test	SF	C	SC	Homologs	Code	SR	Result
1	SFI	**I**	A	1b33/OmpIP	OG5_128023	0.00	NO
				1b17/OMF	OG5_133733	0.00	
			B	1b18/OMA	OG5_155026	0.00	
				1b22/Secretin	OG5_138540	0.00	
			C	1b42/LPS-EP	OG5_140166	0.00	
		III		1b39/OmpW	OG5_138797	0.00	
		III		1b6/OOP	OG5_139592	0.00	
		V		1b9/FadL	OG5_140163	0.00	
		VIII		1b14/OMR	OG5_153441	0.00	
		XIII		1b8/MPP	OG5_127746	0.85	
2	SFI	**I**	A	1b33/OmpIP	OG5_128023	0.45	NO
				1b17/OMF	OG5_133733	0.58	
			B	1b18/OMA	OG5_155026	0.00	
				1b22/Secretin	OG5_138540	0.00	
			C	1b42/LPS-EP	OG5_140166	0.00	
3	SFI	I		1b17/OMF	OG5_133733	0.75	NO
			B	1b18/OMA	OG5_155026	0.00	
				1b22/Secretin	OG5_138540	0.00	
4	SFI	III		1b39/OmpW	OG5_138797	1.00	YES
				1b6/OOP	OG5_139592	1.00	
5	SFIV			1b30/OEP16	OG5_141660	1.00	YES
				1b69/PxMP4	OG5_130976	1.00	

SF = superfamily, C = cluster, SC = subcluster, SR = Sequence retention = the proportion of sequences that pass homology assessment using in PhyloToL. There were 5 tests of homology. The first test evaluates homology in the whole SFI. Test 2 evaluates homology in cluster I of the SFI. Test 3 evaluates homology in the subcluster B of the cluster I of the SFI. Test 4 evaluates homology in the cluster III of the SFI and test 5 evaluates homology in the SFIV. Only the test 4 and 5 show clear evidence of homology with GUIDANCE v2.02 parameters sequence and column cutoff 0.3 and 0.4, respectively.

Table S6. Statistical comparison of Fungi + others root against previously published roots using t-student test.

dataset	H1	H2	t	df	p-value
SEL-	Fungi	Opistho	-2.1541	196.94	0.03245
SEL-	Fungi	Unikonta	-34.607	152.14	< 2.2e-16
SEL-	Fungi	Discoba	-26.327	153.2	< 2.2e-16
SEL-	Fungi	Ancy+Meta	-34.837	147.84	< 2.2e-16
SEL+	Fungi	Opistho	2.5636	196.57	0.01111
SEL+	Fungi	Unikonta	-24.378	138.26	< 2.2e-16
SEL+	Fungi	Discoba	-14.292	177.48	< 2.2e-16
SEL+	Fungi	Ancy+Meta	-23.354	184.33	< 2.2e-16
RAN+	Fungi	Opistho	-1.0961	194.48	0.2744
RAN+	Fungi	Unikonta	-41.909	166.18	< 2.2e-16
RAN+	Fungi	Discoba	-8.113	185.13	6.65E-14
RAN+	Fungi	Ancy+Meta	-22.863	190.89	< 2.2e-16
RAN-	Fungi	Opistho	-11.636	194.45	< 2.2e-16
RAN-	Fungi	Unikonta	-61.562	176.41	2.20E-16
RAN-	Fungi	Discoba	-27.788	187.37	< 2.2e-16
RAN-	Fungi	Ancy+Meta	-45.486	173.9	< 2.2e-16

APPENDIX B

SUPPLEMENTARY FIGURES

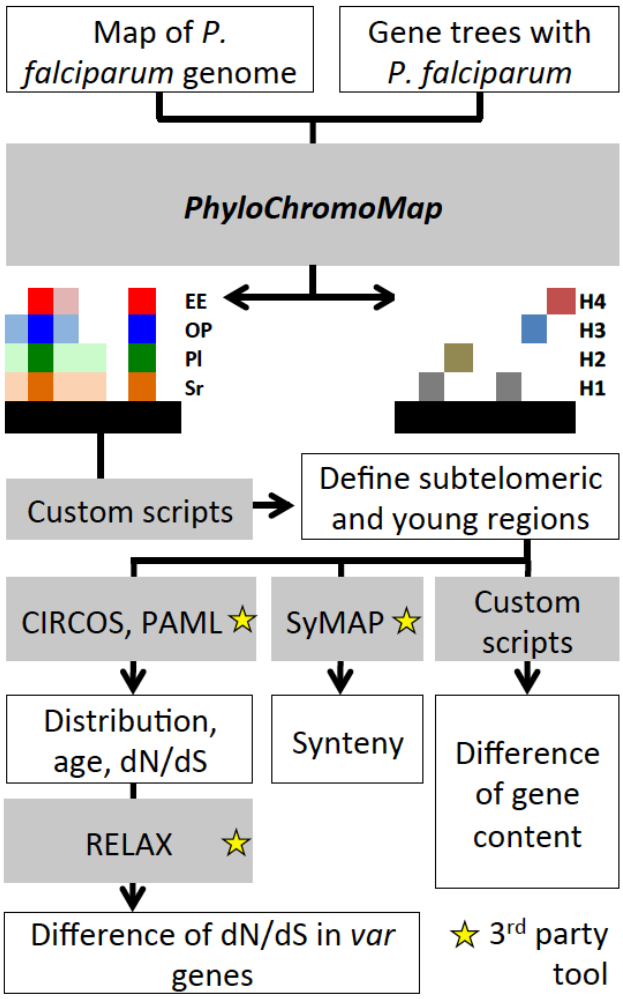

Figure S1. Flow diagram of the methods for mapping the chromosomes of *P. falciparum* with PhyloChromoMap. The genome of *P. falciparum* was compared by BLAST to the database of the Katz lab phylogenomic pipeline in order to build a collection of homologs of the genes of *P. falciparum* that we could then map to chromosomes. We ran these genes through the pipeline (Grant and Katz 2014a; Katz and Grant 2015) to produce a collection of gene trees, which were used by PhyloChromoMap to draw a map of the phylogenetic history of every gene and another map that shows the putative origin of genes based on hypotheses of conservation. We used the resulting phylogenomic map to define the subtelomeric regions based on their relative age (absence of conserved genes). Then, we compared subtelomeric and internal chromosomal regions through analyzes of synteny (using SyMAP), age and dN/dS of paralogs (using CIRCOS and PAML), and difference of gene content (using custom R and phyton scripts). Given that a substantial part of the difference of gene content between subtelomeric and internal regions are due to the antigenic genes, we compared patterns of selection among chromosomal regions in gene families *var*, *rif* and *stevor* using RELAX. However, we present these analyses only for the *var* gene family due to the low number of genes *rif* and absence of *stevor* in intergenic region.

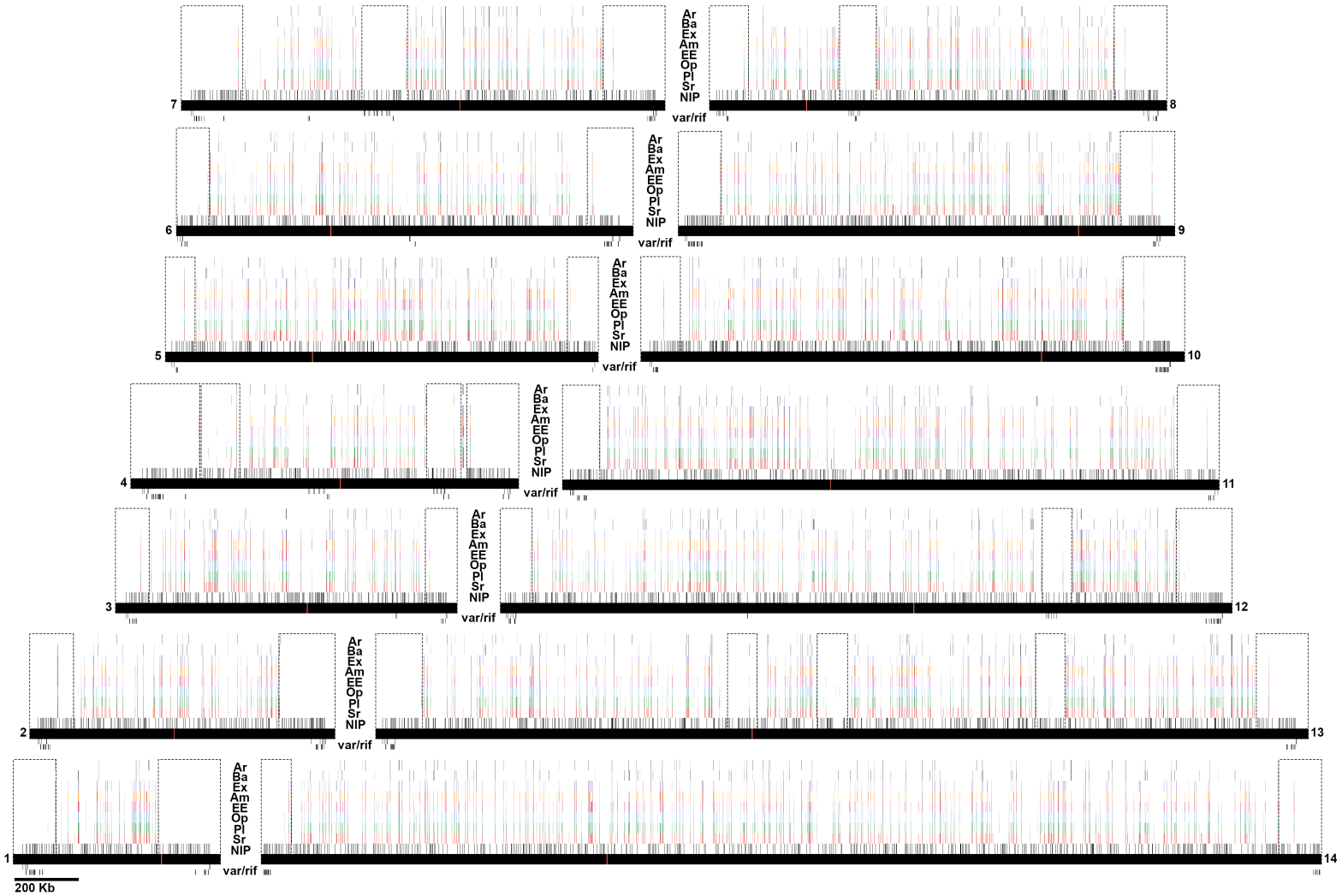

Figure S2. Phylogenomic map of chromosomes of *Plasmodium falciparum 3D7* showing the conservation level of genes assessed. Notes as in Figure 1.1

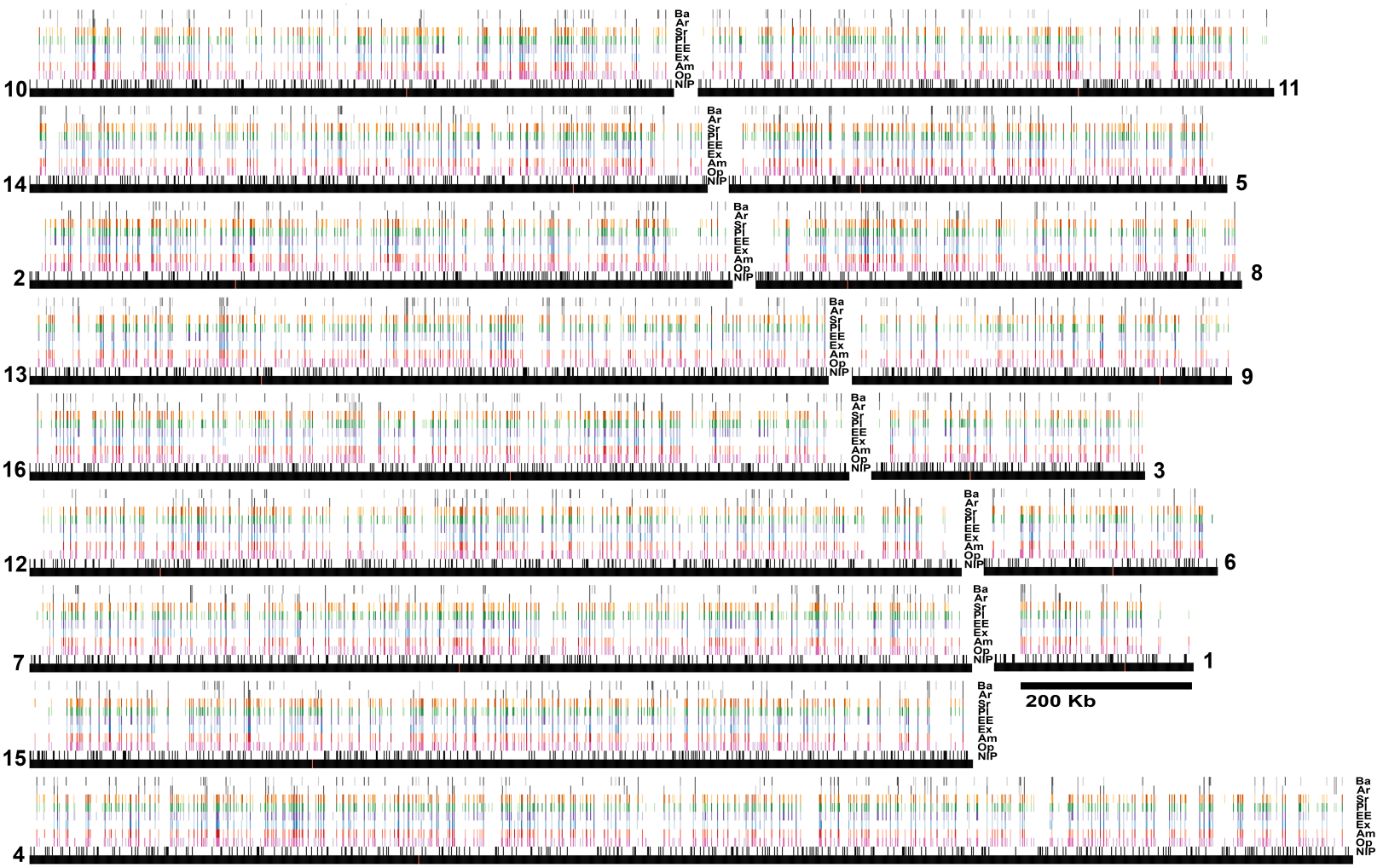

Figure S3. Phylogenomic map of chromosomes of *Saccharomyces cerevisiae* S288C showing the conservation level of genes assessed. Notes as in Figure 1.2

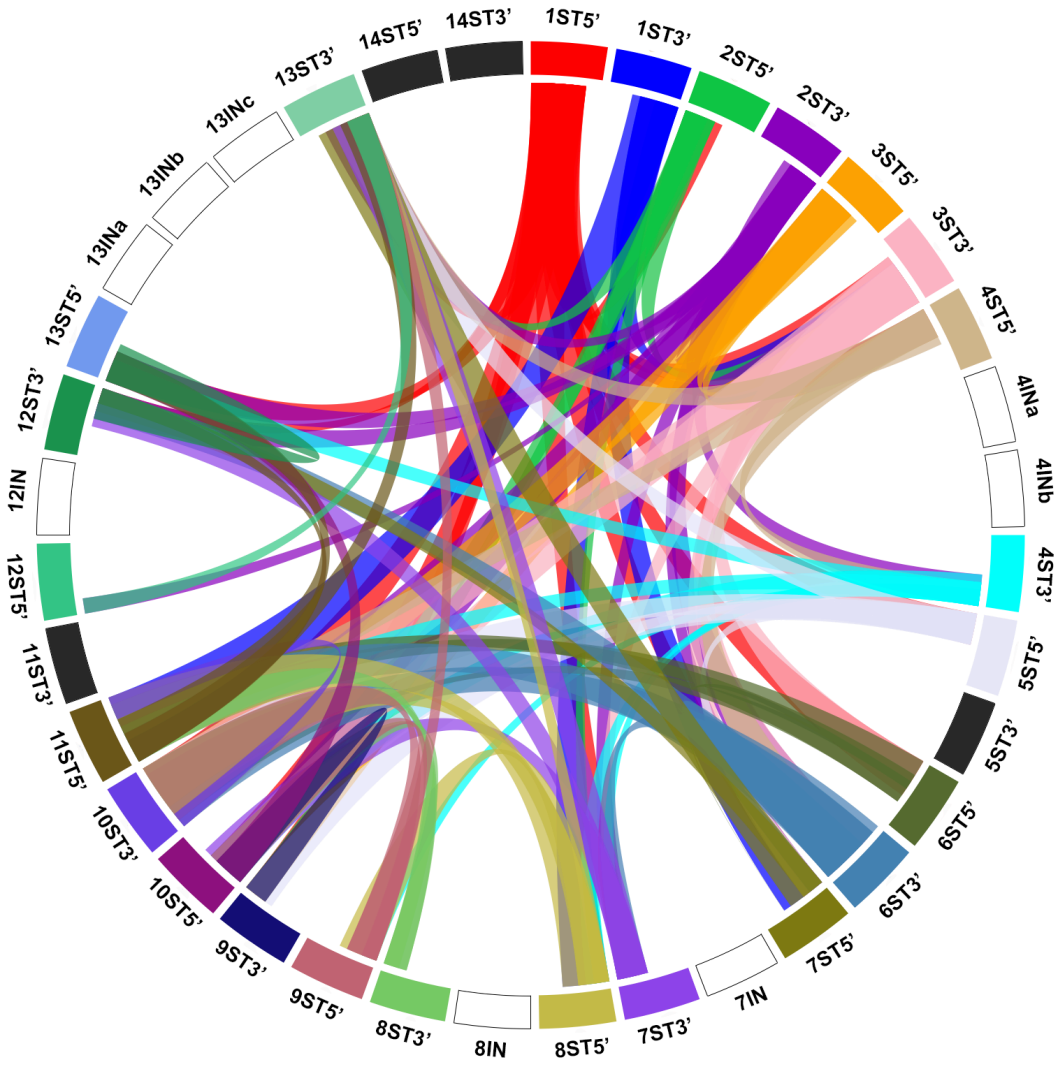

Figure S4. Analysis of synteny shows that synteny blocks are not shared between internal young regions (white boxes) and subtelomeric young regions. Each young region is identified with chromosomal number and chromosomal region (ST for subtelomeric and IN for internal). Subtelomeric young regions are also identified by the chromosomal orientation (5' or 3'). When there is more than one internal young region per chromosome, each region is identified by a letter (e.g. 4INa, 4INb). The colors indicate the synteny blocks shared among young regions and the thickness of the links represents the size of the synteny block. Black and white boxes are young regions (subtelomeric and internal, respectively) that do not share synteny blocks. The size of the box does not represent size of the young region.

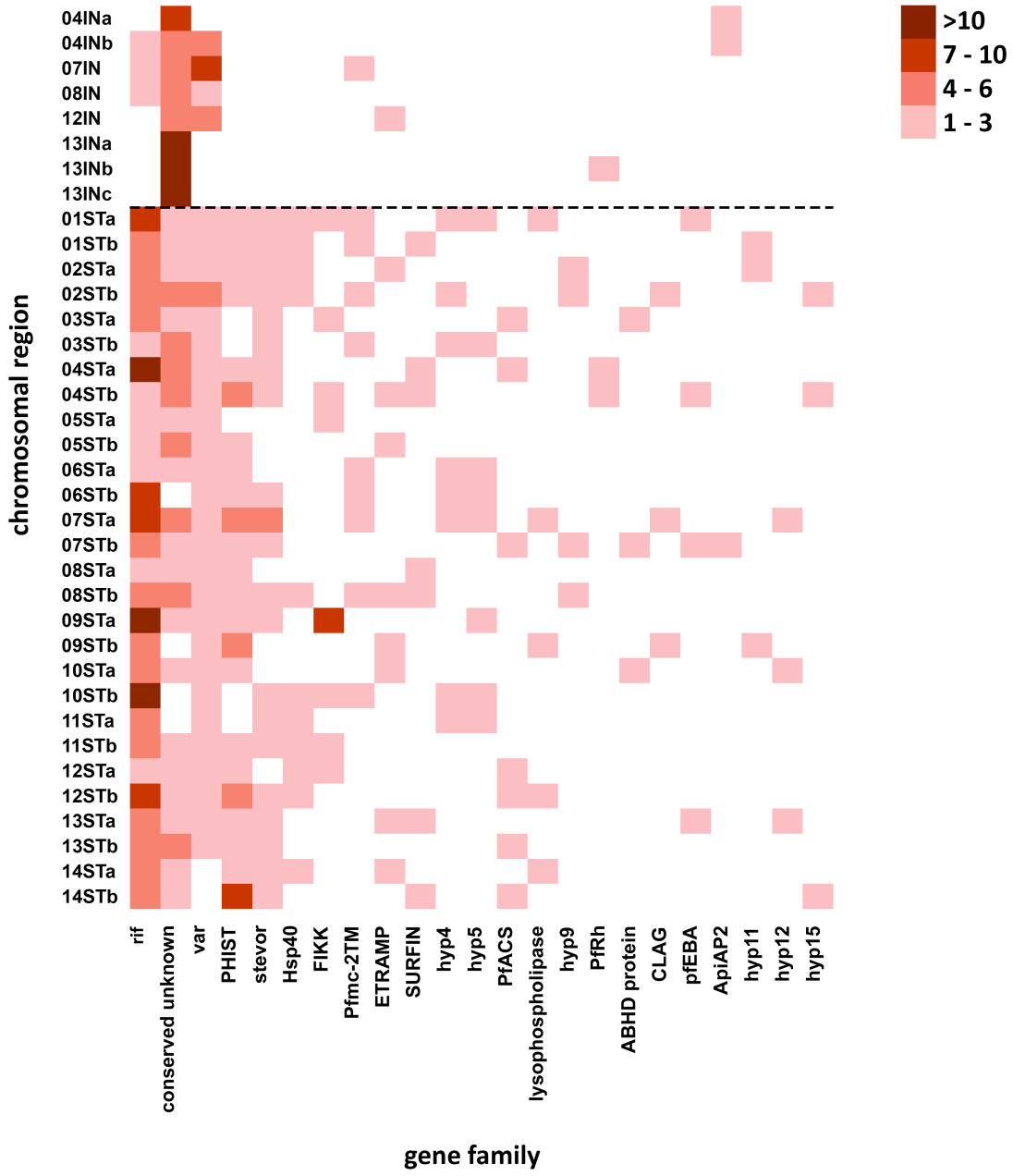

Figure S5. Genes in young regions tend to be restricted to either subtelomeric or internal regions, with the exception of *var* genes that are abundant in both subtelomeric and young regions. This graph is a heatmap of the presence of the proteins or gene families listed on the 'x' axis across the young regions listed on the 'y' axis. Dashed line indicate break between internal (IN) and subtelomeric (ST) regions.

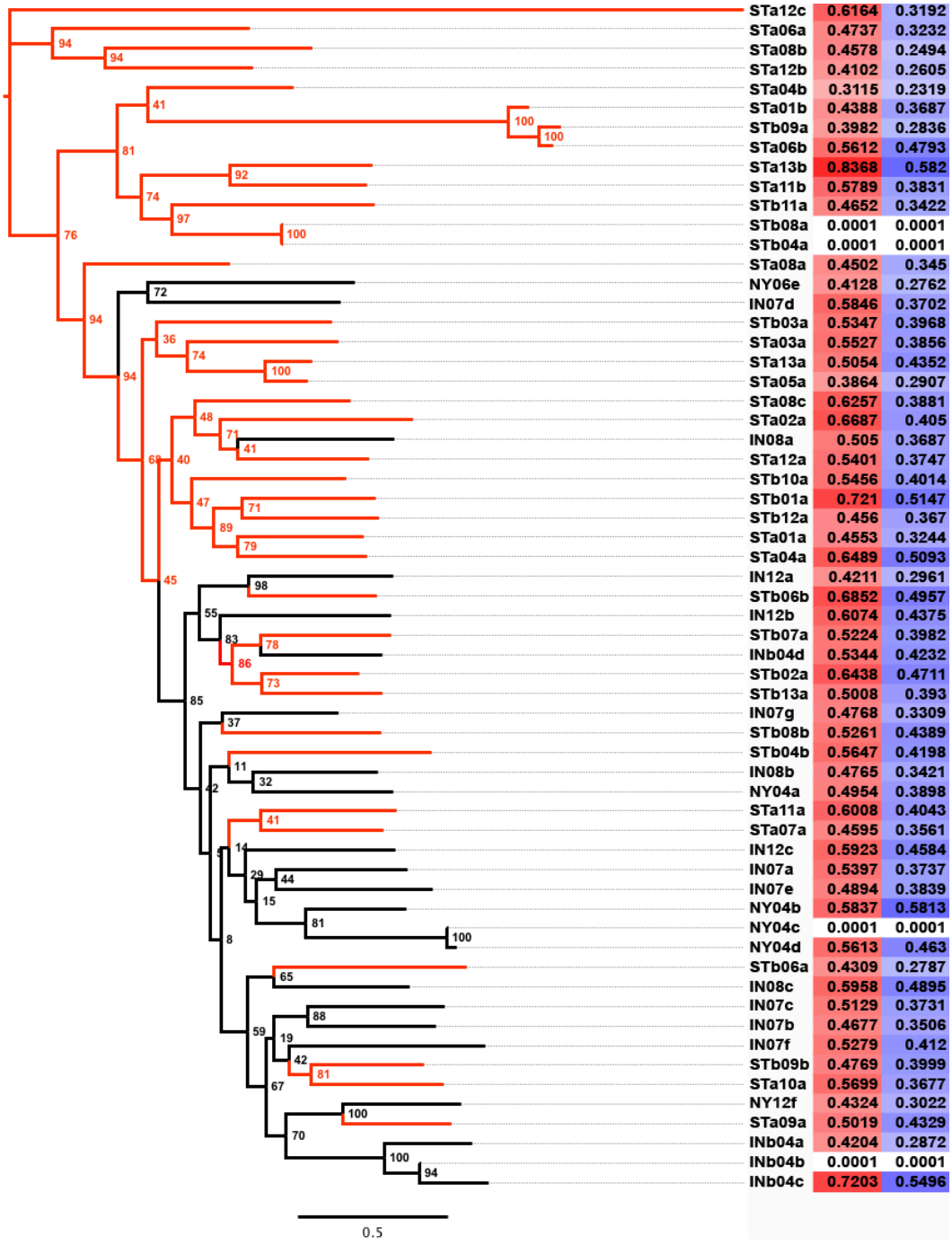

Figure S6. Subtelomeric and internal paralogs of *var* genes do not have significant differences in their dN/dS ratios. Subtelomeric paralogs are represented as red branches and internal paralogs as black branches. Values of dN/dS were calculated with the free ratio model of codeML-PAML(Yang 1997) (red) and HyPhy(Kosakovsky Pond, et al. 2005) (blue). In both cases the darker the color the higher the dN/dS value. The intensity of selection was not significantly different between subtelomeric and internal paralogs (RELAX, k = 1.22, p > 0,05).

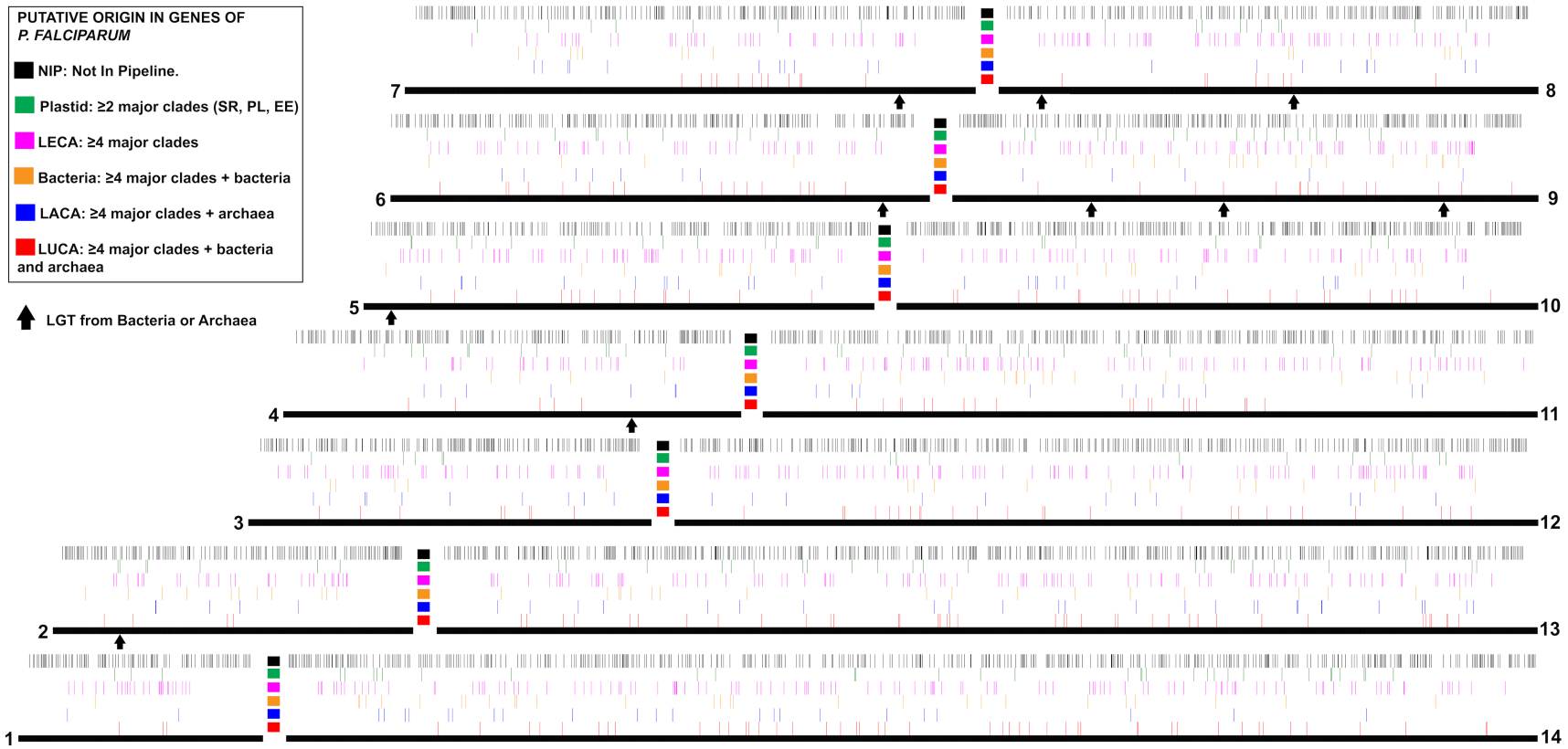

Figure S7. Phylogenomic map of the chromosomes of *P. falciparum* according to the hypothetical origin of genes. Notes as in Figure 1.5.

Figure S8. Detailed phylogenomic map of the chromosomes of *Trypanosoma brucei* generated by combining outputs of PhyloToL with PhyloChromoMap (Cerón-Romero, et al. 2018). Thick black lines represent chromosomes and bars reflect levels of conservation. First row from the bottom (NIP, "not in pipeline") indicates ORFs that do not match our criteria for tree building (i.e. likely *Trypanosoma*-specific, highly divergent or misannotated ORFs). The remaining rows (bottom to top) reflect the presence or absence of the gene in the major clades Excavata (Ex), orphans (EE, "everything else"), Archaeplastida (Pl), SAR (Sr), Amoebozoa (Am), Opisthokonta (Op), Archaea (Ar), and Bacteria (Ba).

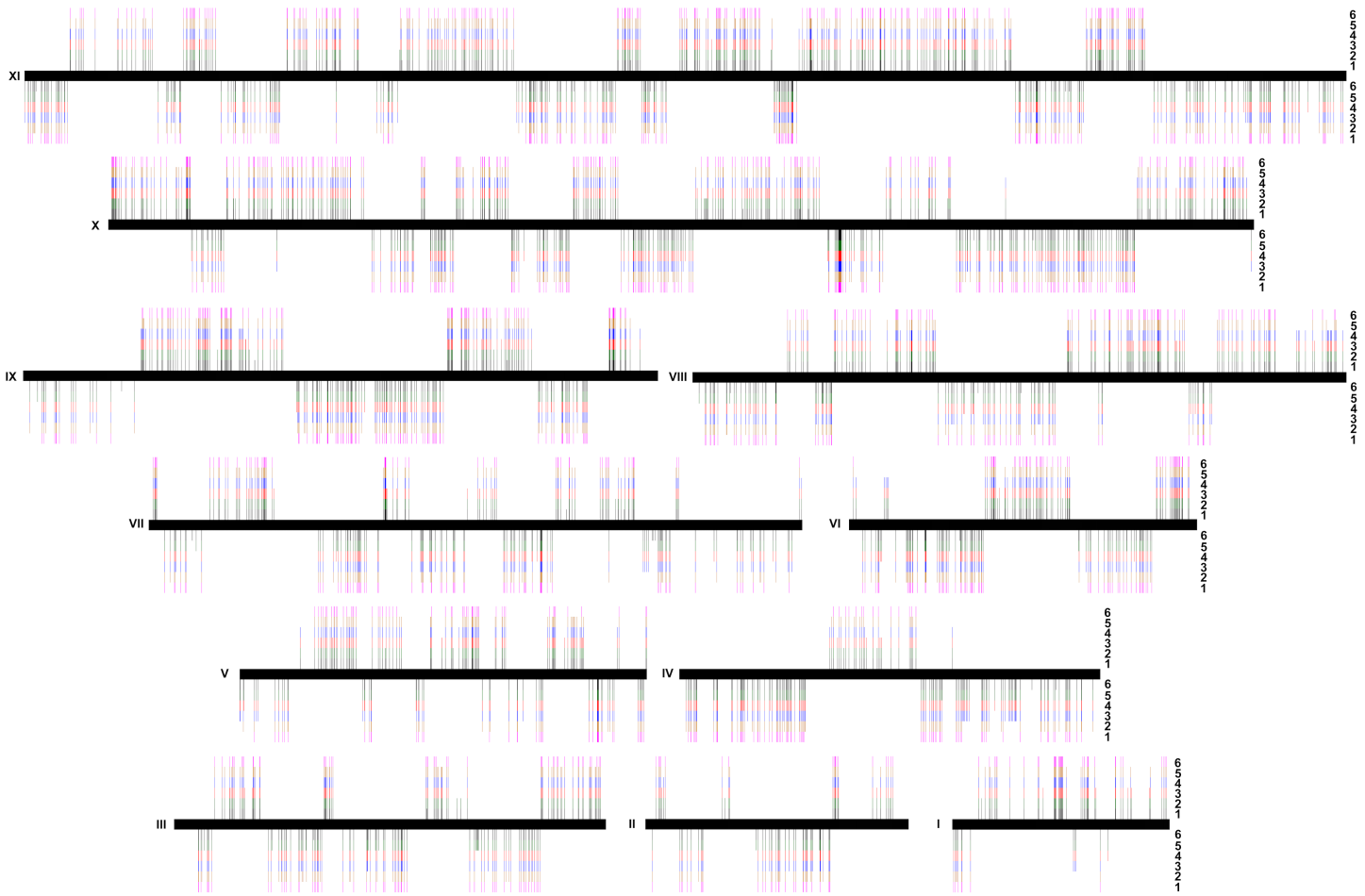

Figure S9. High levels of conservation of many genes across chromosomes (thick lines) of *Trypanosoma brucei*. The height of each bar represents the number of eukaryotic major clades that share the gene, varying from 1-6 major clades (including orphan lineages, EE).

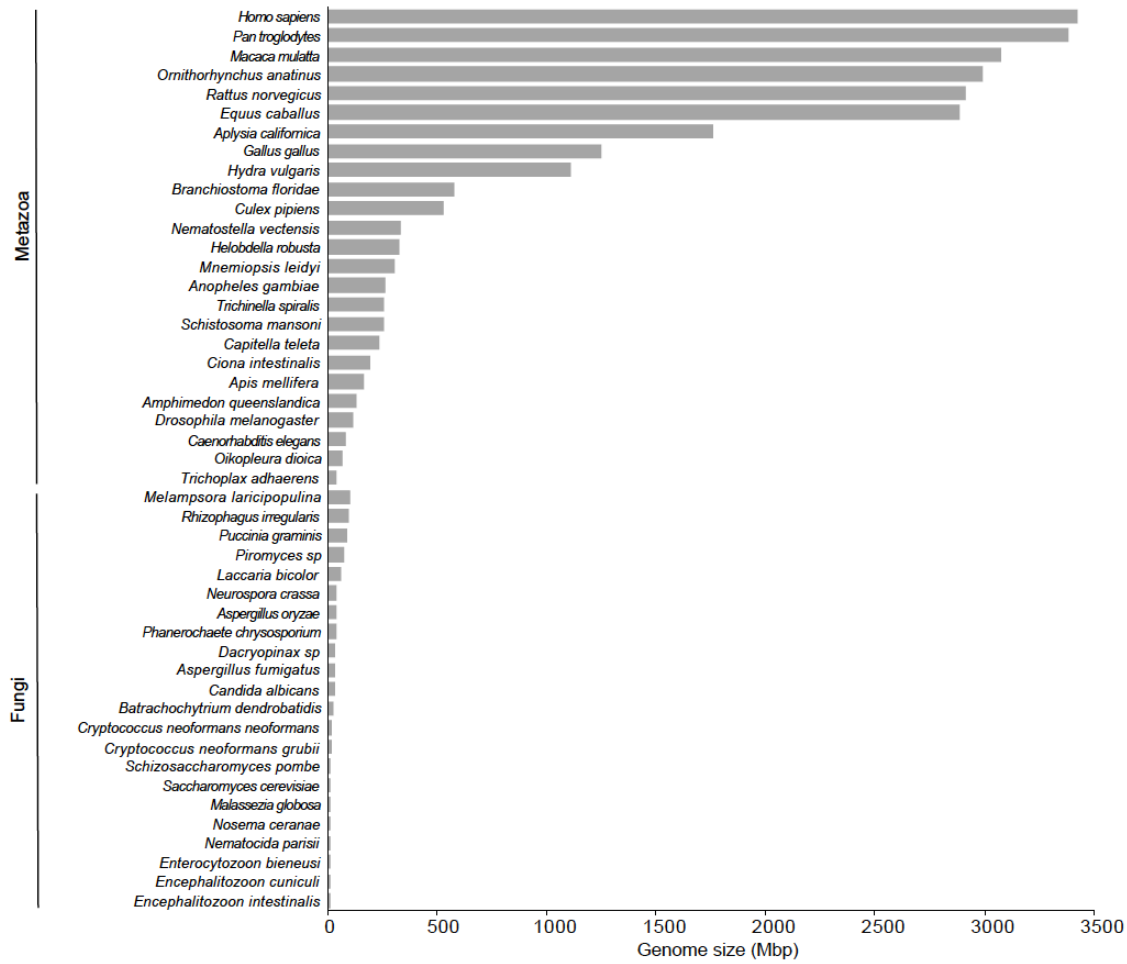

Figure S10. Genome size comparison between the Metazoa and Fungi. The data used for these taxa were whole genome sequences. The fungi genome sizes were taken from JGI (https://jgi.doe.gov/) and the metazoan genome sizes were taken from the Animal Genome Size Database, Release 2.0 (http://www.genomesize.com).

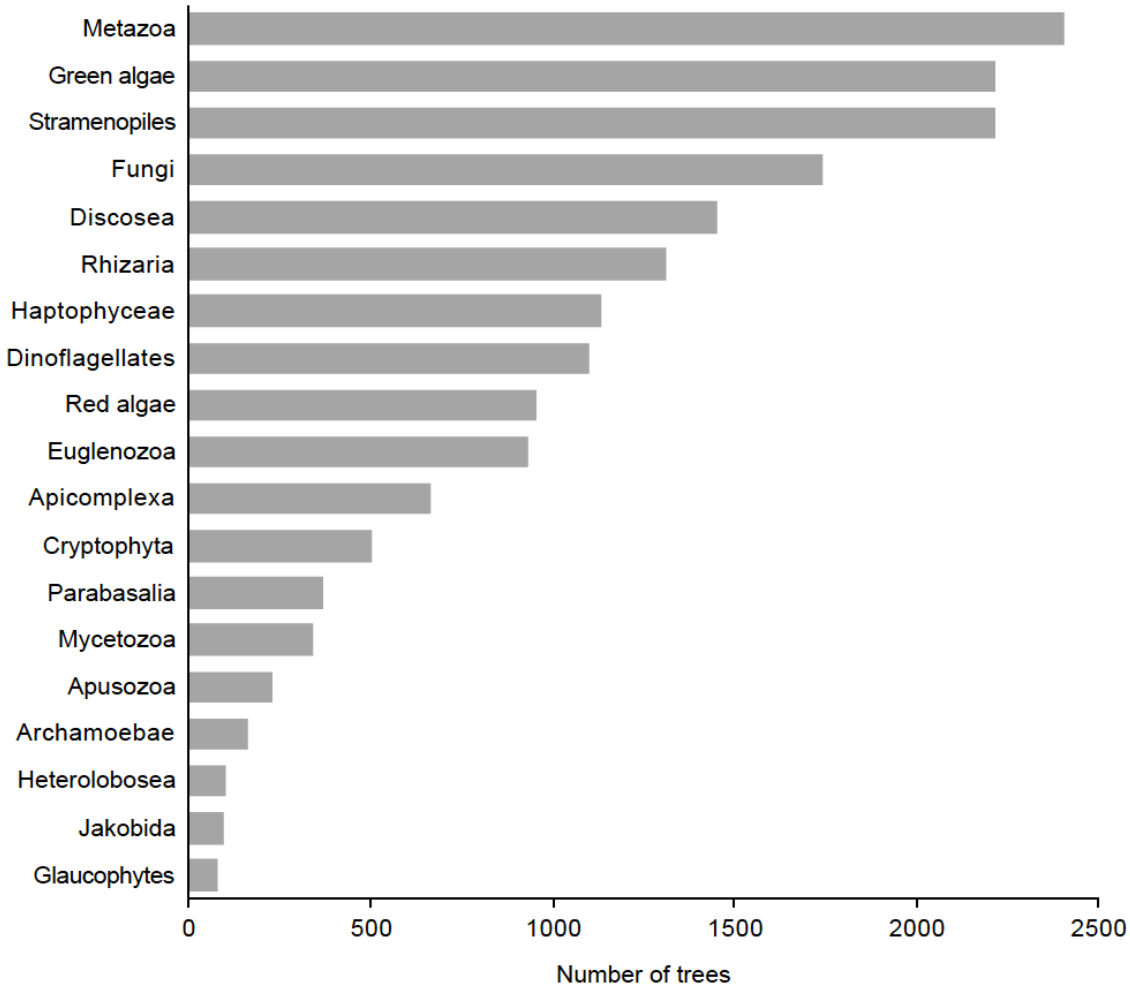

Figure S11. Number of trees with at least three species per minor clade in dataset SEL+. The data used for all these clades were a combination of whole genome sequences and transcriptomes. For Glaucophytes, the most underrepresented clade in the trees, all data came from transcriptomes.

BIBLIOGRAPHY

Adl SM, Simpson AGB, Lane CE, Lukes J, Bass D, Bowser SS, Brown MW, Burki F, Dunthorn M, Hampl V, et al. 2012. The Revised Classification of Eukaryotes. *Journal of Eukaryotic Microbiology* 59:429-493.

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

Ankarklev J, Franzen O, Peirasmaki D, Jerlstrom-Hultqvist J, Lebbad M, Andersson J, Andersson B, Svard SG. 2015. Comparative genomic analyses of freshly isolated Giardia intestinalis assemblage A isolates. *BMC Genomics* 16:697.

Antonelli A, Hettling H, Condamine FL, Vos K, Nilsson RH, Sanderson MJ, Sauquet H, Scharn R, Silvestro D, Topel M, et al. 2017. Toward a Self-Updating Platform for Estimating Rates of Speciation and Migration, Ages, and Relationships of Taxa. *Syst Biol.* 66:152-166.

Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, et al. 2000. Gene Ontology: tool for the unification of biology. *Nat Genet.* 25:25-29.

Bairoch A, Apweiler R. 2000. The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. *Nucleic Acids Res.* 28:45-48.

Baker W, van den Broek A, Camon E, Hingamp P, Sterk P, Stoesser G, Tuli MA. 2000. The EMBL nucleotide sequence database. *Nucleic Acids Res.* 28:19-23.

Baldauf SL, Palmer JD. 1993. Animals and fungi are each other's closest relatives: congruent evidence from multiple proteins. *Proc Natl Acad Sci U S A* 90:11558-11562.

Benson DA, Cavanaugh M, Clark K, Karsch-Mizrachi I, Lipman DJ, Ostell J, Sayers EW. 2017. GenBank. *Nucleic Acids Res.* 45:D37-D42.

Bentley SD, Chater KF, Cerdeno-Tarraga AM, Challis GL, Thomson NR, James KD, Harris DE, Quail MA, Kieser H, Harper D, et al. 2002. Complete genome sequence of the model actinomycete Streptomyces coelicolor A3(2). *Nature* 417:141-147.

Berriman M, Ghedin E, Hertz-Fowler C, Blandin G, Renauld H, Bartholomeu DC, Lennard NJ, Caler E, Hamlin NE, Haas B, et al. 2005. The genome of the African trypanosome *Trypanosoma brucei*. *Science* 309:416-422.

Biderre C, Mathis A, Deplazes P, Weber R, Metenier G, Vivares CP. 1999. Molecular karyotype diversity in the microsporidian *Encephalitozoon cuniculi*. *Parasitology* 118:439-445.

Bowman S, Lawson D, Basham D, Brown D, Chillingworth T, Churcher CM, Craig A, Davies RM, Devlin K, Feltwell T, et al. 1999. The complete nucleotide sequence of chromosome 3 of Plasmodium falciparum. *Nature* 400:532-538.

Brown MW, Heiss AA, Kamikawa R, Inagaki Y, Yabuki A, Tice AK, Shiratori T, Ishida KI, Hashimoto T, Simpson AGB, et al. 2018. Phylogenomics places orphan protistan lineages in a novel eukaryotic super-group. *Genome Biol Evol.* 10:427-433.

Burge C, Karlin S. 1997. Prediction of complete gene structures in human genomic DNA. *J Mol Biol.* 268:78-94.

Burki F, Imanian B, Hehenberger E, Hirakawa Y, Maruyama S, Keeling PJ. 2014. Endosymbiotic gene transfer in tertiary plastid-containing dinoflagellates. *Eukaryot Cell* 13:246-255.

Burki F, Kaplan M, Tikhonenkov DV, Zlatogursky V, Minh BQ, Radaykina LV, Smirnov A, Mylnikov AP, Keeling PJ. 2016. Untangling the early diversification of eukaryotes: a phylogenomic study of the evolutionary origins of Centrohelida, Haptophyta and Cryptista. *Proc Biol Sci.* 283:20152802.

Burki F, Shalchian-Tabrizi K, Minge M, Skjaeveland A, Nikolaev SI, Jakobsen KS, Pawlowski J. 2007. Phylogenomics reshuffles the eukaryotic supergroups. *PLOS ONE* 2:e790.

Burleigh JG, Bansal MS, Eulenstein O, Hartmann S, Wehe A, Vision TJ. 2011. Genome-scale phylogenetics: inferring the plant tree of life from 18,896 gene trees. *Syst Biol.* 60:117-125.

Bussey H, Kaback DB, Zhong W, Vo DT, Clark MW, Fortin N, Hall J, Ouellette BF, Keng T, Barton AB, et al. 1995. The nucleotide sequence of chromosome I from Saccharomyces cerevisiae. *Proc Natl Acad Sci U S A* 92:3809-3813.

Butterfield NJ. 2009. Modes of pre-Ediacaran multicellularity. *Precambrian Res* 173:201-211.

Butterfield NJ. 2005. Probable Proterozoic fungi. *Paleobiology* 31:165-182.

Cantarel BL, Korf I, Robb SM, Parra G, Ross E, Moore B, Holt C, Sanchez Alvarado A, Yandell M. 2008. MAKER: an easy-to-use annotation pipeline designed for emerging model organism genomes. *Genome Res.* 18:188-196.

Carlton JM, Adams JH, Silva JC, Bidwell SL, Lorenzi H, Caler E, Crabtree J, Angiuoli SV, Merino EF, Amedeo P, et al. 2008. Comparative genomics of the neglected human malaria parasite Plasmodium vivax. *Nature* 455:757-763.

Carlton JM, Angiuoli SV, Suh BB, Kooij TW, Pertea M, Silva JC, Ermolaeva MD, Allen JE, Selengut JD, Koo HL, et al. 2002. Genome sequence and comparative analysis of the model rodent malaria parasite Plasmodium yoelii yoelii. *Nature* 419:512-519.

Carlton JM, Galinski MR, Barnwell JW, Dame JB. 1999. Karyotype and synteny among the chromosomes of all four species of human malaria parasite. *Mol Biochem Parasitol* 101:23-32.

Caron F, Meyer E. 1985. Does *Paramecium primaurelia* use a different genetic code in its macronucleus? *Nature* 314:185-188.

Cavalier-Smith T. (lower eukaryotes co-authors). 1993. Kingdom Protozoa and its 18 phyla. *Micro. Rev.* 57:953-994.

Cavalier-Smith T. 1989. Molecular phylogeny. Archaebacteria and Archezoa. *Nature* 339:l00-01.

Cavalier-Smith T, Fiore-Donno AM, Chao E, Kudryavtsev A, Berney C, Snell EA, Lewis R. 2015. Multigene phylogeny resolves deep branching of Amoebozoa. *Mol Phylogenet Evol* 83:293-304.

Ceron-Romero MA, Maurer-Alcala XX, Grattepanche JD, Yan Y, Fonseca MM, Katz LA. 2019. PhyloToL: A Taxon/Gene-Rich Phylogenomic Pipeline to Explore Genome Evolution of Diverse Eukaryotes. *Mol Biol Evol* 36:1831-1842.

Ceron-Romero MA, Nwaka E, Owoade Z, Katz LA. 2018. PhyloChromoMap, a Tool for Mapping Phylogenomic History along Chromosomes, Reveals the Dynamic Nature of Karyotype Evolution in Plasmodium falciparum. *Genome Biol Evol* 10:553-561.

Cerón-Romero MA, Nwaka E, Owoade Z, Katz LA. 2018. PhyloChromoMap, a tool for mapping phylogenomic history along chromosomes, reveals the dynamic nature of karyotype evolution in *Plasmodium falciparum*. *Genome Biol Evol.* 10:553-561.

Chater KF. 2016. Recent advances in understanding Streptomyces. *F1000Res* 5:2795.

Chaudhary R, Bansal MS, Wehe A, Fernandez-Baca D, Eulenstein O. 2010. iGTP: A software package for large-scale gene tree parsimony analysis. *BMC Bioinformatics* 11.

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

Claessens A, Hamilton WL, Kekre M, Otto TD, Faizullabhoy A, Rayner JC, Kwiatkowski D. 2014. Generation of antigenic diversity in Plasmodium falciparum by structured rearrangement of Var genes during mitosis. *PLoS Genet* 10:e1004812.

Daniels JP, Gull K, Wickstead B. 2010. Cell biology of the *Trypanosome* genome. *Microbiol Mol Biol Rev.* 74:552-569.

Darriba D, Taboada GL, Doallo R, Posada D. 2011. ProtTest 3: fast selection of best-fit models of protein evolution. *Bioinformatics* 27:1164-1165.

Davis WJ, Amses KR, Benny GL, Carter-House D, Chang Y, Grigoriev I, Smith ME, Spatafora JW, Stajich JE, James TY. 2019. Genome-scale phylogenetics reveals a monophyletic Zoopagales (Zoopagomycota, Fungi). *Mol Phylogenet Evol* 133:152-163.

de Bruin D, Lanzer M, Ravetch JV. 1994. The polymorphic subtelomeric regions of Plasmodium falciparum chromosomes contain arrays of repetitive sequence elements. *Proc Natl Acad Sci U S A* 91:619-623.

de Duve C. 2007. The origin of eukaryotes: a reappraisal. *Nat Rev Genet* 8:395-403.

De Oliveira Martins L, Mallo D, Posada D. 2016. A Bayesian Supertree Model for Genome-Wide Species Tree Reconstruction. *Syst Biol* 65:397-416.

Dehal P, Boore JL. 2005. Two rounds of whole genome duplication in the ancestral vertebrate. *PLoS Biol* 3:e314.

Delarbre S, Gatti S, Scaglia M, Drancourt M. 2001. Genetic diversity in the microsporidian *Encephalitozoon hellem* demonstrated by pulsed-field gel electrophoresis. *Journal of Eukaryotic Microbiology* 48:471-474.

Derelle R, Lang FB. 2011. Rooting the eukaryotic tree with mitochondrial and bacterial proteins. *Mol Biol Evol*.

Derelle R, Torruella G, Klimes V, Brinkmann H, Kim E, Vlcek C, Lang BF, Elias M. 2015. Bacterial proteins pinpoint a single eukaryotic root. *Proceedings of the National Academy of Sciences of the United States of America* 112:E693-E699.

Dia N, Lavie L, Faye N, Metenier G, Yeramian E, Duroure C, Toguebaye BS, Frutos R, Niang MN, Vivares CP, et al. 2016. Subtelomere organization in the genome of the microsporidian Encephalitozoon cuniculi: patterns of repeated sequences and physicochemical signatures. *BMC Genomics* 17:34.

dos Reis M, Inoue J, Hasegawa M, Asher RJ, Donoghue PCJ, Yang ZH. 2012. Phylogenomic datasets provide both precision and accuracy in estimating the timescale of placental mammal phylogeny. *Proc Biol Sci.* 279:3491-3500.

Drori M, Rice A, Einhorn M, Chay O, Glick L, Mayrose I. 2018. OneTwoTree: An online tool for phylogeny reconstruction. *Mol Ecol Resour.* 18:1492-1499.

Dunn CW, Hejnol A, Matus DQ, Pang K, Browne WE, Smith SA, Seaver E, Rouse GW, Obst M, Edgecombe GD, et al. 2008. Broad phylogenomic sampling improves resolution of the animal tree of life. *Nature* 452:745-749.

Edgar RC. 2010. Search and clustering orders of magnitude faster than BLAST. *Bioinformatics* 26:2460-2461.

El-Sayed NM, Myler PJ, Blandin G, Berriman M, Crabtree J, Aggarwal G, Caler E, Renauld H, Worthey EA, Hertz-Fowler C, et al. 2005. Comparative genomics of trypanosomatid parasitic protozoa. *Science* 309:404-409.

Embley TM, Hirt RP. (Embley co-authors). 1998. Early branching eukaryotes? *Current Opinion in Genectics & Development* 8:624-629.

Figueiredo L, Scherf A. 2005. Plasmodium telomeres and telomerase: the usual actors in an unusual scenario. *Chromosome Res* 13:517-524.

Figueiredo LM, Freitas-Junior LH, Bottius E, Olivo-Marin JC, Scherf A. 2002. A central role for Plasmodium falciparum subtelomeric regions in spatial positioning and telomere length regulation. *EMBO J* 21:815-824.

Figueiredo LM, Pirrit LA, Scherf A. 2000. Genomic organisation and chromatin structure of Plasmodium falciparum chromosome ends. *Mol Biochem Parasitol* 106:169-174.

Foster PG. 2004. Modeling compositional heterogeneity. *Syst Biol.* 53:485-495.

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

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

Galtier N, Daubin V. 2008. Dealing with incongruence in phylogenomic analyses. *Philosophical Transactions of the Royal Society B-Biological Sciences* 363:4023-4029.

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

Gardner MJ, Tettelin H, Carucci DJ, Cummings LM, Aravind L, Koonin EV, Shallom S, Mason T, Yu K, Fujii C, et al. (99021743 co-authors). 1998. Chromosome 2 sequence of the human malaria parasite Plasmodium falciparum. *Science* 282:1126-1132.

Ghedin E, Bringaud F, Peterson J, Myler P, Berriman M, Ivens A, Andersson B, Bontempi E, Eisen J, Angiuoli S, et al. 2004. Gene synteny and evolution of genome architecture in trypanosomatids. *Mol Biochem Parasitol* 134:183-191.

Grant JR, Katz LA. 2014a. Building a phylogenomic pipeline for the eukaryotic tree of life - addressing deep phylogenies with genome-scale data. *PLoS Curr.* 6.

Grant JR, Katz LA. 2014b. Phylogenomic study indicates widespread lateral gene transfer in *Entamoeba* and suggests a past intimate relationship with parabasalids. *Genome Biol Evol.* 6:2350-2360.

Guigo R, Muchnik I, Smith TF. 1996. Reconstruction of ancient molecular phylogeny. *Molecular Phylogenetics and Evolution* 6:189-213.

Hahn MW. 2009. Distinguishing Among Evolutionary Models for the Maintenance of Gene Duplicates. *Journal of Heredity* 100:605-617.

Hall BG. 2013. Building phylogenetic trees from molecular data with MEGA. *Mol Biol Evol.* 30:1229-1235.

Hall N, Pain A, Berriman M, Churcher C, Harris B, Harris D, Mungall K, Bowman S, Atkin R, Baker S, et al. 2002. Sequence of Plasmodium falciparum chromosomes 1, 3-9 and 13. *Nature* 419:527-531.

Hampl V, Hug L, Leigh JW, Dacks JB, Lang BF, Simpson AG, Roger AJ. 2009. Phylogenomic analyses support the monophyly of Excavata and resolve relationships among eukaryotic "supergroups". *Proc Natl Acad Sci U S A* 106:3859-3864.

He D, Fiz-Palacios O, Fu CJ, Fehling J, Tsai CC, Baldauf SL. 2014. An Alternative Root for the Eukaryote Tree of Life. *Curr Biol.* 24:465-470.

Heaphy SM, Mariotti M, Gladyshev VN, Atkins JF, Baranov PV. 2016. Novel ciliate genetic code variants including the reassignment of all three stop codons to sense codons in *Condylostoma magnum*. *Mol Biol Evol.* 33:2885-2889.

Heiss AA, Kolisko M, Ekelund F, Brown MW, Roger AJ, Simpson AGB. 2018. Combined morphological and phylogenomic re-examination of malawimonads, a critical taxon for inferring the evolutionary history of eukaryotes. *R Soc Open Sci.* 5:171707.

Hernandez-Rivas R, Herrera-Solorio AM, Sierra-Miranda M, Delgadillo DM, Vargas M. 2013. Impact of chromosome ends on the biology and virulence of *Plasmodium falciparum*. *Mol Biochem Parasitol.* 187:121-128.

Hernandez-Rivas R, Perez-Toledo K, Herrera Solorio AM, Delgadillo DM, Vargas M. 2010. Telomeric heterochromatin in Plasmodium falciparum. *J Biomed Biotechnol* 2010:290501.

Hirt RP, Logsdon JM, Healy B, Dorey MW, Doolittle WF, Embley TM. (S: RPB1 co-authors). 1999. Microsporidia are related to Fungi: Evidence from the largest subunit of RNA polymerase II and other proteins. *Proceedings of the National Academy of Sciences of the United States of America* 96:580-585.

Hope RM. 1993. Selected Features of Marsupial Genetics. *Genetica* 90:165-180.

Huelsenbeck JP, Ronquist F. 2001. MRBAYES: Bayesian inference of phylogenetic trees. *Bioinformatics* 17:754-755.

Hug LA, Baker BJ, Anantharaman K, Brown CT, Probst AJ, Castelle CJ, Butterfield CN, Hernsdorf AW, Amano Y, Ise K, et al. 2016. A new view of the tree of life. *Nat Microbiol.* 1:16048.

Ikeda H, Ishikawa J, Hanamoto A, Shinose M, Kikuchi H, Shiba T, Sakaki Y, Hattori M, Omura S. 2003. Complete genome sequence and comparative analysis of the industrial microorganism Streptomyces avermitilis. *Nat Biotechnol* 21:526-531.

Jackson CJ, Reyes-Prieto A. 2014. The mitochondrial genomes of the glaucophytes Gloeochaete wittrockiana and Cyanoptyche gloeocystis: multilocus phylogenetics suggests a monophyletic archaeplastida. *Genome Biol Evol* 6:2774-2785.

Kanehisa M, Goto S. 2000. KEGG: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.* 28:27-30.

Katoh K, Standley DM. 2013. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol Biol Evol.* 30:772-780.

Katz LA. 2012. Origin and diversification of eukaryotes. *Annu Rev Microbiol* 66:411-427.

Katz LA. 2015. Recent events dominate interdomain lateral gene transfers between prokaryotes and eukaryotes and, with the exception of endosymbiotic gene transfers, few ancient transfer events persist. *Philosophical Transactions of the Royal Society B-Biological Sciences* 370.

Katz LA, Grant JR. 2015. Taxon-rich phylogenomic analyses resolve the eukaryotic tree of life and reveal the power of subsampling by sites. *Syst Biol.* 64:406-415.

Katz LA, Grant JR, Parfrey LW, Burleigh JG. 2012. Turning the crown upside down: gene tree parsimony roots the eukaryotic tree of life. *Syst Biol.*

Keeling PJ, Doolittle WF. 1997. Evidence that eukaryotic triosephosphate isomerase is of alpha-proteobacterial origin. *Proc Natl Acad Sci U S A.* 94:1270-1275.

Keeling PJ, Leander BS. 2003. Characterisation of a non-canonical genetic code in the oxymonad *Streblomastix strix*. *J Mol Biol.* 326:1337-1349.

Kellis M, Patterson N, Endrizzi M, Birren B, Lander ES. 2003. Sequencing and comparison of yeast species to identify genes and regulatory elements. *Nature* 423:241-254.

Kissinger JC, DeBarry J. 2011. Genome cartography: charting the apicomplexan genome. *Trends Parasitol* 27:345-354.

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

Korf I. 2004. Gene finding in novel genomes. *BMC Bioinformatics* 5:59.

Kosakovsky Pond SL, Frost SDW, Muse SV. 2005. HyPhy: hypothesis testing using phylogenies. *Bioinformatics* 21:676-679.

Kryukov K, Imanishi T. 2016. Human contamination in public genome assemblies. *PLoS One* 11:e0162424.

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

Kumar S, Krabberod AK, Neumann RS, Michalickova K, Zhao S, Zhang X, Shalchian-Tabrizi K. 2015. BIR pipeline for preparation of phylogenomic data. *Evol Bioinform Online.* 11:79-83.

Kuo CH, Wares JP, Kissinger JC. 2008. The Apicomplexan whole-genome phylogeny: an analysis of incongruence among gene trees. *Mol Biol Evol* 25:2689-2698.

Kyes SA, Kraemer SM, Smith JD. 2007. Antigenic variation in Plasmodium falciparum: gene organization and regulation of the var multigene family. *Eukaryot Cell* 6:1511-1520.

Lane CE, Archibald JM. 2006. Novel nucleomorph genome architecture in the cryptomonad genus hemiselmis. *Journal of Eukaryotic Microbiology* 53:515-521.

Lane CE, Khan H, MacKinnon M, Fong A, Theophilou S, Archibald JM. 2006. Proceedings of the SMBE Tri-National Young Investigators' Workshop 2005. Insight into the diversity and evolution of the cryptomonad nucleomorph genome. *Mol Biol Evol.* 23:856-865.

Langer MR, Lipps JH. 1995. Phylogenetic incongruence between dinoflagellate endosymbionts (*Symbiodinium*) and their host foraminifera (*Sorites*): Small-subunit ribosomal RNA gene sequence evidence. *Mar Micropaleontol.* 26:179-186.

Larsson AJM, Stanley G, Sinha R, Weissman IL, Sandberg R. 2018. Computational correction of index switching in multiplexed sequencing libraries. *Nat Methods.* 15:305-307.

Laurin-Lemay S, Brinkmann H, Philippe H. 2012. Origin of land plants revisited in the light of sequence contamination and missing data. *Curr Biol.* 22:R593-594.

Leigh JW, Susko E, Baumgartner M, Roger AJ. 2008. Testing congruence in phylogenomic analysis. *Syst Biol.* 57:104-115.

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

Loidl J, Nairz K. 1997. Karyotype variability in yeast caused by nonallelic recombination in haploid meiosis. *Genetics* 146:79-88.

Loron CC, Francois C, Rainbird RH, Turner EC, Borensztajn S, Javaux EJ. 2019. Early fungi from the Proterozoic era in Arctic Canada. *Nature* 570:232-235.

Lozupone CA, Knight RD, Landweber LF. 2001. The molecular basis of nuclear genetic code change in ciliates. *Curr Biol.* 11:65-74.

Lusk RW. 2014. Diverse and widespread contamination evident in the unmapped depths of high throughput sequencing data. *PLoS One* 9:e110808.

Maddison WP. 1997. Gene trees in species trees. *Syst Biol.* 46:523-536.

Mallo D, Posada D. 2016. Multilocus inference of species trees and DNA barcoding. *Philos Trans R Soc Lond B Biol Sci.* 371:20150335.

Martin W, Rotte C, Hoffmeister M, Theissen U, Gelius-Dietrich G, Ahr S, Henze K. 2003. Early cell evolution, eukaryotes, anoxia, sulfide, oxygen, fungi first (?), and a tree of genomes revisited. *IUBMB Life* 55:193-204.

Martinez-Calvillo S, Vizuet-de-Rueda JC, Florencio-Martinez LE, Manning-Cela RG, Figueroa-Angulo EE. 2010. Gene expression in trypanosomatid parasites. *J Biomed Biotechnol.* 2010:525241.

McGrath CL, Katz LA. 2004. Genome diversity in microbial eukaryotes. *Trends Ecol. Evol.* 19:32-38.

Merchant S, Wood DE, Salzberg SL. 2014. Unexpected cross-species contamination in genome sequencing projects. *PeerJ* 2:e675.

Mirarab S, Reaz R, Bayzid MS, Zimmermann T, Swenson MS, Warnow T. 2014. ASTRAL: genome-scale coalescent-based species tree estimation. *Bioinformatics* 30:i541-548.

Mirarab S, Warnow T. 2015. ASTRAL-II: coalescent-based species tree estimation with many hundreds of taxa and thousands of genes. *Bioinformatics* 31:i44-52.

Mitra A, Skrzypczak M, Ginalski K, Rowicka M. 2015. Strategies for achieving high sequencing accuracy for low diversity samples and avoiding sample bleeding using illumina platform. *PLoS One* 10:e0120520.

Narechania A, Baker RH, Sit R, Kolokotronis SO, DeSalle R, Planet PJ. 2012. Random Addition Concatenation Analysis: a novel approach to the exploration of phylogenomic

signal reveals strong agreement between core and shell genomic partitions in the cyanobacteria. *Genome Biol Evol.* 4:30-43.

Okamoto N, Chantangsi C, Horak A, Leander BS, Keeling PJ. 2009. Molecular Phylogeny and Description of the Novel Katablepharid *Roombia truncata* gen. et sp nov., and Establishment of the Hacrobia Taxon nov. *PLoS One* 4:e7080.

Oliverio AM, Katz LA. 2014. The dynamic nature of genomes across the tree of life. *Genome Biol Evol.* 6:482-488.

Otto SP, Whitton J. 2000. Polyploid incidence and evolution. *Annu Rev Genet* 34:401-437.

Pace T, Ponzi M, Scotti R, Frontali C. 1995. Structure and superstructure of Plasmodium falciparum subtelomeric regions. *Mol Biochem Parasitol* 69:257-268.

Pain A, Bohme U, Berry AE, Mungall K, Finn RD, Jackson AP, Mourier T, Mistry J, Pasini EM, Aslett MA, et al. 2008. The genome of the simian and human malaria parasite Plasmodium knowlesi. *Nature* 455:799-803.

Palenik B, Grimwood J, Aerts A, Rouze P, Salamov A, Putnam N, Dupont C, Jorgensen R, Derelle E, Rombauts S, et al. 2007. The tiny eukaryote Ostreococcus provides genomic insights into the paradox of plankton speciation. *Proceedings of the National Academy of Sciences of the United States of America* 104:7705-7710.

Panchy N, Lehti-Shiu M, Shiu SH. 2016. Evolution of Gene Duplication in Plants. *Plant Physiol* 171:2294-2316.

Panek T, Zihala D, Sokol M, Derelle R, Klimes V, Hradilova M, Zadrobilkova E, Susko E, Roger AJ, Cepicka I, et al. 2017. Nuclear genetic codes with a different meaning of the UAG and the UAA codon. *BMC Biol.* 15:8.

Parfrey LW, Lahr DJG, Katz LA. 2008. The dynamic nature of eukaryotic genomes. *Mol Biol Evol.* 25:787-794.

Parfrey LW, Lahr DJG, Knoll AH, Katz LA. 2011. Estimating the timing of early eukaryotic diversification with multigene molecular clocks. *Proceedings of the National Academy of Sciences of the United States of America* 108:13624-13629.

Parks DH, Imelfort M, Skennerton CT, Hugenholtz P, Tyson GW. 2015. CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. *Genome Res.* 25:1043-1055.

Penn O, Privman E, Ashkenazy H, Landan G, Graur D, Pupko T. 2010. GUIDANCE: a web server for assessing alignment confidence scores. *Nucleic Acids Res.* 38:W23-W28.

Philippe H, Vienne D, Ranwez V, Roure B, Baurain D, Delsuc F. 2017. Pitfalls in supermatrix phylogenomics. *Eur J Tax.* 283:1–25.

Privman E, Penn O, Pupko T. 2012. Improving the Performance of Positive Selection Inference by Filtering Unreliable Alignment Regions. *Mol Biol Evol.* 29:1-5.

Reddy BL, Saier MH, Jr. 2016. Properties and Phylogeny of 76 Families of Bacterial and Eukaryotic Organellar Outer Membrane Pore-Forming Proteins. *PLoS One* 11:e0152733.

Rice P, Longden I, Bleasby A. 2000. EMBOSS: The European molecular biology open software suite. *Trends Genet.* 16:276-277.

Rodriguez-Ezpeleta N, Brinkmann H, Burey SC, Roure B, Burger G, Loffelhardt W, Bohnert HJ, Philippe H, Lang BF. 2005. Monophyly of primary photosynthetic eukaryotes: Green plants, red algae, and glaucophytes. *Curr Biol.* 15:1325-1330.

Roger AJ. 1999. Reconstructing early events in eukaryotic evolution. *Am. Nat.* 154:S146-S163.

Rogozin IB, Basu MK, Csuros M, Koonin EV. 2009. Analysis of rare genomic changes does not support the unikont-bikont phylogeny and suggests cyanobacterial symbiosis as the point of primary radiation of eukaryotes. *Genome Biol Evol* 1:99-113.

Ronquist F, Teslenko M, van der Mark P, Ayres DL, Darling A, Hohna S, Larget B, Liu L, Suchard MA, Huelsenbeck JP. 2012. MrBayes 3.2: efficient Bayesian phylogenetic inference and model choice across a large model space. *Syst Biol* 61:539-542.

Sanderson MJ, Boss D, Chen D, Cranston KA, Wehe A. 2008. The PhyLoTA browser: Processing GenBank for molecular phylogenetics research. *Syst Biol.* 57:335-346.

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

Scherf A, Lopez-Rubio JJ, Riviere L. 2008. Antigenic variation in Plasmodium falciparum. *Annu Rev Microbiol* 62:445-470.

Schnoes AM, Brown SD, Dodevski I, Babbitt PC. 2009. Annotation error in public databases: misannotation of molecular function in enzyme superfamilies. *PLoS Comput Biol.* 5:e1000605.

Schubert I, Vu GTH. 2016. Genome Stability and Evolution: Attempting a Holistic View. *Trends in Plant Science* 21:749-757.

Sela I, Ashkenazy H, Katoh K, Pupko T. 2015a. GUIDANCE2: accurate detection of unreliable alignment regions accounting for the uncertainty of multiple parameters. *Nucleic Acids Res* 43:W7-14.

Sela I, Ashkenazy H, Katoh K, Pupko T. 2015b. GUIDANCE2: accurate detection of unreliable alignment regions accounting for the uncertainty of multiple parameters. *Nucleic Acids Res.* 43:W7-14.

Seligy VL, James AP. 1977. Multiplicity and distribution of rDNA cistrons among chromosome I and VII aneuploids of Saccharomyces cerevisiae. *Exp Cell Res* 105:63-72.

Shrestha PM, Nevin KP, Shrestha M, Lovley DR. 2013. When Is a Microbial Culture "Pure"? Persistent Cryptic Contaminant Escapes Detection Even with Deep Genome Sequencing. *Mbio.* 4:e00591-00512.

Silver TD, Moore CE, Archibald JM. 2010. Nucleomorph ribosomal DNA and telomere dynamics in chlorarachniophyte algae. *J Eukaryot Microbiol* 57:453-459.

Sites JW, Reed KM. 1994. Chromosomal Evolution, Speciation, and Systematics - Some Relevant Issues. *Herpetologica* 50:237-249.

Smith SA, Beaulieu JM, Donoghue MJ. 2009. Mega-phylogeny approach for comparative biology: an alternative to supertree and supermatrix approaches. *BMC Evol Biol.* 9:37.

Soderlund C, Nelson W, Shoemaker A, Paterson A. 2006. SyMAP: A system for discovering and viewing syntenic regions of FPC maps. *Genome Res* 16:1159-1168.

Stahl HD, Crewther PE, Anders RF, Kemp DJ. 1987. Structure of the FIRA gene of Plasmodium falciparum. *Mol Biol Med* 4:199-211.

Stamatakis A. 2014. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* 30:1312-1313.

Stamatakis A, Ott M, Ludwig T. 2005. RAxML-OMP: An efficient program for phylogenetic inference on SMPs. *Lecture Notes Computer Sci.* 3606:288-302.

Stanke M, Morgenstern B. 2005. AUGUSTUS: a web server for gene prediction in eukaryotes that allows user-defined constraints. *Nucleic Acids Res.* 33:W465-467.

Stechmann A, Cavalier-Smith T. 2003. The root of the eukaryote tree pinpointed. *Curr Biol* 13:R665-666.

Stechmann A, Cavalier-Smith T. 2002. Rooting the eukaryote tree by using a derived gene fusion. *Science* 297:89-91.

Steenkamp ET, Wright J, Baldauf SL. 2006. The protistan origins of animals and fungi. *Mol Biol Evol* 23:93-106.

Struck TH. 2013. The impact of paralogy on phylogenomic studies - a case study on annelid relationships. *PLoS One* 8:e62892.

Su XZ, Heatwole VM, Wertheimer SP, Guinet F, Herrfeldt JA, Peterson DS, Ravetch JA, Wellems TE. 1995. The large diverse gene family var encodes proteins involved in cytoadherence and antigenic variation of Plasmodium falciparum-infected erythrocytes. *Cell* 82:89-100.

Sugiura M, Tanaka Y, Suzaki T, Harumoto T. 2012. Alternative gene expression in type I and type II cells may enable further nuclear changes during conjugation of *Blepharisma japonicum*. *Protist* 163:204-216.

Swart EC, Serra V, Petroni G, Nowacki M. 2016. Genetic codes with no dedicated stop codon: context-dependent translation termination. *Cell* 166:691-702.

Tanifuji G, Kim E, Onodera NT, Gibeault R, Dlutek M, Cawthorn RJ, Fiala I, Lukes J, Greenwood SJ, Archibald JM. 2011. Genomic Characterization of *Neoparamoeba pemaquidensis* (Amoebozoa) and Its Kinetoplastid Endosymbiont. *Eukaryot Cell.* 10:1143-1146.

Tanifuji G, Onodera NT, Brown MW, Curtis BA, Roger AJ, Ka-Shu Wong G, Melkonian M, Archibald JM. 2014. Nucleomorph and plastid genome sequences of the chlorarachniophyte Lotharella oceanica: convergent reductive evolution and frequent recombination in nucleomorph-bearing algae. *BMC Genomics* 15:374.

Team RC. 2016. R: A language and environment for statistical computing [Internet]. Vienna, Austria.

Tremblay-Savard O, Swenson KM. 2012. A graph-theoretic approach for inparalog detection. *BMC Bioinformatics* 13 Suppl 19:S16.

Van de Peer Y, Ben Ali A, Meyer A. 2000. Microsporidia: accumulating molecular evidence that a group of amitochondriate and suspectedly primitive eukaryotes are just curious fungi. *Gene* 246:1-8.

Vasilakis N, Forrester NL, Palacios G, Nasar F, Savji N, Rossi SL, Guzman H, Wood TG, Popov V, Gorchakov R, et al. 2013. Negevirus: a proposed new taxon of insect-specific viruses with wide geographic distribution. *J Virol.* 87:2475-2488.

Vinuesa P, Ochoa-Sanchez LE, Contreras-Moreira B. 2018. GET_PHYLOMARKERS, a Software Package to Select Optimal Orthologous Clusters for Phylogenomics and Inferring Pan-Genome Phylogenies, Used for a Critical Geno-Taxonomic Revision of the Genus *Stenotrophomonas*. *Front Microbiol.* 9:771.

Walther A, Hesselbart A, Wendland J. 2014. Genome sequence of Saccharomyces carlsbergensis, the world's first pure culture lager yeast. *G3 (Bethesda)* 4:783-793.

Wertheim JO, Murrell B, Smith MD, Kosakovsky Pond SL, Scheffler K. 2015. RELAX: detecting relaxed selection in a phylogenetic framework. *Mol Biol Evol* 32:820-832.

Whidden C, Zeh N, Beiko RG. 2014. Supertrees Based on the Subtree Prune-and-Regraft Distance. *Syst Biol* 63:566-581.

Wickett NJ, Mirarab S, Nguyen N, Warnow T, Carpenter E, Matasci N, Ayyampalayam S, Barker MS, Burleigh JG, Gitzendanner MA, et al. 2014. Phylotranscriptomic analysis

of the origin and early diversification of land plants. *Proc Natl Acad Sci U S A.* 111:E4859-4868.

Wideman JG, Gawryluk RM, Gray MW, Dacks JB. 2013. The ancient and widespread nature of the ER-mitochondria encounter structure. *Mol Biol Evol* 30:2044-2049.

Williams TA, Foster PG, Cox CJ, Embley TM. 2013. An archaeal origin of eukaryotes supports only two primary domains of life. *Nature* 504:231-236.

Wolfe KHaDCS. (fungi co-authors). 1997. Molecular Evidence for an Ancient Duplication of the Entire Yeast Genome. *Nature* 387:708-713.

Wu M, Eisen JA. 2008. A simple, fast, and accurate method of phylogenomic inference. *Genome Biol.* 9:R151.

Yang Z. 1997. PAML: a program package for phylogenetic analysis by maximum likelihood. *Comput Appl Biosci* 13:555-556.

Zufall RA, Robinson T, Katz LA. 2005. Evolution of developmentally regulated genome rearrangements in eukaryotes. *Journal of Experimental Zoology Part B-Molecular and Developmental Evolution* 304B:448-455.