1-19-2019

PhyloChromoMap, a Tool for Mapping Phylogenomic History along Chromosomes, Reveals the Dynamic Nature of Karyotype Evolution in Plasmodium falciparum

Mario A. Cerón-Romero
Smith College

Esther Nwaka
Smith College

Zuliat Owoade
Smith College

Laura A. Katz
Smith College, lkatz@smith.edu

Follow this and additional works at: https://scholarworks.smith.edu/bio_facpubs

Part of the Biology Commons

Recommended Citation
https://scholarworks.smith.edu/bio_facpubs/48

This Article has been accepted for inclusion in Biological Sciences: Faculty Publications by an authorized administrator of Smith ScholarWorks. For more information, please contact scholarworks@smith.edu
**PhyloChromoMap, a Tool for Mapping Phylogenomic History along Chromosomes, Reveals the Dynamic Nature of Karyotype Evolution in *Plasmodium falciparum***

Mario A. Cerón-Romero1,2, Esther Nwaka1, Zuliat Owoade1, and Laura A. Katz1,2,*

1Department of Biological Sciences, Smith College, Northampton, Massachusetts
2Program in Organismic and Evolutionary Biology, University of Massachusetts Amherst

*Corresponding author: E-mail: lkatz@smith.edu.

**Abstract**

The genome of *Plasmodium 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 2,116 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 with 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 whereas older paralogs are enriched in internal regions. Combining these observations with analyses of synteny, we demonstrate that subtelomeric regions 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 versus 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.

**Key words:** chromosomal mapping, *Plasmodium falciparum*, phylogenomics, karyotype evolution, antigenic genes.

**Introduction**

Numerous studies of plants, animals, and fungi have informed the 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 the last decades have 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 nonhomologous chromosomes (i.e., ectopic recombination) can lead to intra-specific 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), *Encephalitozoon hellem* (Delarbre et al. 2001), and *Plasmodium falciparum* (Freitas-Junior et al. 2000; 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. 2000; Delarbre et al. 2001).

The growing number of genomes that are available enables the development of new methods to explore patterns of karyotype evolution. Well-annotated genomes can be used to build physical maps in order to compare structural...
characteristics such as 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), and Trypanosoma (Ghedin et al. 2004). Likewise for phylogenomic analyses, the increase in genomic data provides more taxa and genes to compare. Analysis of the phylogenetic history of genes along chromosome combining these maps 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, 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 (that 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,” containing a series of repeat sequences (Figueiredo et al. 2000, 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. 2000; 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, 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).

In this study we 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 2014; 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 [Stramenopiles, Alveolata, Rhizaria], 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.

Materials and Methods

Development of PhyloChromoMap

Starting from a phylogenomic pipeline previously built in our lab (Grant and Katz 2014; Katz and Grant 2015), we develop PhyloChromoMap to map the evolutionary history of genes along chromosomes (https://github.com/Katzlab/PhyloChromoMap.py; last accessed January 2018). Our initial collection of homologs uses gene families defined in OrthoMCL (http://www.orthomcl.org/orthomcl/, last accessed January 2018) and as such, each of these clusters of homologs is referred to as an “orthologous group” or OG. We analyze a total of 5,336 putative coding genes from P. falciparum 3D7 (assembly ASM276v1) by BLAST (Altschul et al. 1990) against OrthoMCL (supplementary fig. S1, Supplementary Material online). This results in 2,116 genes falling in 1,962 OGs that are represented in our pipeline. The remaining OGs are not represented in our taxon-rich 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 (fig. 1 and supplementary figs. S1 and S2, Supplementary Material online). 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 implement PhyloChromoMap also in the model organism S. cerevisiae S288C, mapping 3338 of its 5893 ORFs (fig. 2 and supplementary fig. S3, Supplementary Material online).

Definition of Subtelomeres and Detection of Young Portions and Centromeres

We define 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 (supplementary fig. S1, Supplementary Material online). Young portions are defined as regions in which genes are in <3 major eukaryotic clades, though we allow the presence of one gene conserved in three 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; see supplementary Materials,
We search young portions in both subtelomeric and internal regions, only considering internal young portions that are ≥ 90 kb (supplementary table S1, Supplementary Material online). All chromosomes except chromosome 10 have an internal 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 (supplementary table S2, Supplementary Material online).

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; supplementary fig. S1, Supplementary Material online). 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 (supplementary fig. S4, Supplementary Material online). 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.

**FIG. 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 misannotated 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.

**FIG. 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 misannotated 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), and Bacteria (Ba) that contain the indicated gene. Unlike to all the other chromosomes (supplementary fig. S2), chromosome 1 exhibits large regions of low gene content toward the ends.
Finally, overlapping synteny blocks are merged. We also survey the gene content of young portions, including *Plasmodium* specific coding domains (supplementary fig. S1, Supplementary Material online). We categorize the sequences by gene family when possible and plot their frequency as a heatmap (supplementary fig. S5, Supplementary Material online).

We use CIRCOS plots (Krzywinski et al. 2009) to map paralogs of genes that match OGs (fig. 3 and supplementary fig. S1, Supplementary Material online). In CIRCOS, we choose the option “links” 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, pairwise dN/dS values are calculated for all paralogs using yn00, PAML (Yang 1997) and compared between subtelomeric and internal paralogs (fig. 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; supplementary fig. S6, Supplementary Material online). 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.

**Analysis of Putative Origin of Genes**

We use two approaches to detect both recent and old interdomain LGT events in *P. falciparum*, a parametric approach based on nucleotide composition and a phylogenetic approach (supplementary table S3, Supplementary Material online). For the parametric approach, we calculate the average GC content per chromosome and per gene; when the average GC content in a gene is 2 SD 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 the phylogenetic toolkit P4 (Foster 2004). In the topology of the gene trees, we identify potential interdomain LGTs when: (1) the gene trees contain only prokaryotes and Apicomplexa; and (2) 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 6 eukaryotic
major clades (considering orphans ["EE"—everything else] as a major clade) are candidate Endosymbiotic Gene Transfers (EGTs) from mitochondria. Genes that have bacteria and 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 (fig. 5 and supplementary fig. S7, Supplementary Material online).

Results

Development of PhyloChromoMap

We build PhyloChromoMap to map the evolutionary history of genes along chromosomes using P. falciparum as a test case. In sum, we start with a collection of 13,104 multisequence alignments generated in Guidance (Sela et al. 2015) and corresponding gene trees built in RaxML (Stamatakis 2014), which includes up to 519 Eukaryotes, 303 Bacteria and 119 Archaea (Grant and Katz 2014; 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 Materials and Methods, table 1). We then use the function “image” in R (Team 2016) to map the phylogenetic conservation of each gene along each chromosome.

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). 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 misannotated; these genes are considered “NIP” in our analyses as they do not pass our criteria for generation of multisequence alignments and trees (see Materials and Methods, table 1).

We build phylogenomic maps of the 14 chromosomes of P. falciparum 3D7 to illuminate patterns of conservation across different chromosomal regions (fig. 1 and supplementary fig. S2, Supplementary Material online). Distinct patterns of conservation are found across chromosomes. For instance, whereas 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; supplementary table S1, Supplementary Material online), and internal young regions average 106 kb (range of 91–141 kb; supplementary table S1, Supplementary Material online). 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; fig. 1 and supplementary fig. S2, Supplementary Material online).

To exemplify further the power of PhyloChromoMap, we also generate the phylogenic map of the chromosomes of S. cerevisiae in order to validate our method (fig. 2 and supplementary fig. S3, Supplementary Material online). Overall this map shows a higher density of genes than we observe.

Figure 4.—Paralogs from gene family var (blue) do not exhibit significant differences in selection intensity (i.e., dN/dS) according to location, whereas 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 data sets 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 ratios 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; supplementary fig. S6, Supplementary Material online) and are under less functional constraint than most internal paralogs.
Bacteria) are in magenta, blue, and red, respectively. (in at least 4 eukaryotic major clades), LACA (in at least 4 eukaryotic major clades and Archaea), and LUCA (in at least 4 eukaryotic major clades, Archaea and prokaryotes to Apicomplexa. NIP: Not in pipeline, likely young genes, are in black. Candidate EGTs from plastid (in at least 2 photosynthetic major clades [i.e., Sr, Pl, EE]) and mitochondria (in at least 4 eukaryotic major clades and Bacteria) are in green and orange, respectively. Candidate conserved genes from LECA (in at least 2 photosynthetic major clades [i.e., Sr, Pl, EE]) and LECA (in at least 4 eukaryotic major clades and Archaea), and LUCA (in at least 4 eukaryotic major clades, Archaea and Bacteria) are in magenta, blue, and red, respectively.

for *P. falciparum* and here too we do not see any pattern of gene conservation near the centromeres (fig. 2 and supplementary fig. S3, Supplementary Material online). 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 (fig. 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).

**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 (supplementary fig. S4, Supplementary Material online) and comparison of gene content (supplementary fig. S5, Supplementary Material online). Chromosomes share blocks of sequences in conserved order (i.e., synteny blocks) in subtelomeric regions with a few exceptions (14ST3*, 14ST5*, 5ST3*, and 11ST3*; supplementary fig. S4, Supplementary Material online). Some subtelomeric regions (e.g., 13ST3*, 13TS*, 11ST5*) have complex patterns of synteny, sharing many blocks with other subtelomeric regions. In contrast, internal young regions 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* (supplementary fig. S5, Supplementary Material online).

**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 (fig. 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 (fig. 1 and supplementary fig. S5, Supplementary Material online). Moreover, dN/dS ratios are relatively high for *var* genes (mean 0.5, 95% CI 0.46–0.54; fig. 4 and supplementary fig. S6, Supplementary Material online). 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).

**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 (supplementary table S3, Supplementary Material online). This gene (FIRA) is an interspersed repeat
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 with internal chromosome regions, which contain a mixture of conserved and lineage-specific genes (fig. 1 and supplementary fig. S2, Supplementary Material online). These data and the evidence of syntenic blocks among subtelomeres (supplementary fig. S4, Supplementary Material online) are consistent with the hypothesis that chromosomes of *P. falciparum* are actively swapping subtelomeric regions due to frequent ectopic recombination (Freitas-Junior et al. 2000; Scherf et al. 2001, 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. 2000).

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 its 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 *E. 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 as 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).

**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 with those throughout the rest of the chromosome map (fig. 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 frequently during the shuffling of subtelomeric regions between chromosomes, leading to differences

**Discussion**

**Patterns of Gene Conservation in *P. falciparum* and Other Eukaryotes**

Here, we present PhyloChromoMap, a novel method that combines the power of phylogenetics 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 with internal chromosome regions, which contain a mixture of conserved and lineage-specific genes (fig. 1 and supplementary fig. S2, Supplementary Material online). These data and the evidence of syntenic blocks among subtelomeres (supplementary fig. S4, Supplementary Material online) are consistent with the hypothesis that chromosomes of *P. falciparum* are actively swapping subtelomeric regions due to frequent ectopic recombination (Freitas-Junior et al. 2000; Scherf et al. 2001, 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. 2000).

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 its 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 *E. 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 as 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).

**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 with those throughout the rest of the chromosome map (fig. 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 frequently during the shuffling of subtelomeric regions between chromosomes, leading to differences
of gene content between subtelomeric and internal regions in *P. falciparum* (supplementary fig. S5, Supplementary Material online). For instance, subtelomeric regions in *P. falciparum* are enriched for the rapidly evolving immune response gene families such as var, nif, stevar (Freitas-Junior et al. 2000; 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 chromosome maps using dNdS ratios (fig. 4 and supplementary fig. S6, Supplementary Material online). We compare patterns for the var gene family, which are deployed as the parasite seeks to evade host immune responses (Su et al. 1995; Scher et al. 2008; Claessens et al. 2014), to paralogs of other gene families in both subtelomeric and internal regions (fig. 4). Overall, paralogs of subtelomeric gene families are under less selection constraint than paralogs of internal regions as evidenced by dNdS ratios (fig. 4). 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. In contrast, patterns for var paralogs are not affected by their position in the chromosome (fig. 4 and supplementary fig. S6, Supplementary Material online). 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 the location along the chromosome.

**Putative Origin of Each Gene of *P. falciparum***

*PhyloChromoMap* enables exploration of the age and origin of genes along chromosome maps. For example, we identify three candidate LGTs (i.e., 1-cys peroxiredoxin, ribosomal protein L35 precursor and holo-ACP synthase, supplementary table S3, Supplementary Material online) as potential EGTs as they encode for apicoplast functions such as fatty acid synthesis. We can then map cases of EGT and LGT along chromosomes for *P. falciparum* 3D7 (fig. 5 and supplementary fig. S7, Supplementary Material online). We also bind genes into categories based on possible age (fig. 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.

**Supplementary Material**

Supplementary data are available at Genome Biology and Evolution online.


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


