Tail Wags the Dog? Functional Gene Classes Driving Genome-Wide GC Content in Plasmodium spp.

Andreina I. Castillo1, Andrew D. Nelson2, and Eric Lyons3,*

1School of Environmental Science, Policy, and Management, University of California, Berkeley
2School of Plant Sciences, University of Arizona
3BIO5 Institute, School of Plant Sciences, University of Arizona

*Corresponding author: E-mail: ericlyons@email.arizona.edu.

Accepted: January 18, 2019

Data deposition: This project used publicly available data and integrated in the comparative genomics platform, CoGe. No new data were generated and thus we cannot use this statement.

Abstract

Plasmodium parasites are valuable models to understand how nucleotide composition affects mutation, diversification, and adaptation. No other observed eukaryotes have undergone such large changes in genomic Guanine–Cytosine (GC) content as seen in the genus Plasmodium (~30% within 35–40 Myr). Although mutational biases are known to influence GC content in the human-infective Plasmodium vivax and Plasmodium falciparum; no study has addressed how different gene functional classes contribute to genus-wide compositional changes, or if Plasmodium GC content variation is driven by natural selection. Here, we tested the hypothesis that certain gene processes and functions drive variation in global GC content between Plasmodium species. We performed a large-scale comparative genomic analysis using the genomes and predicted genes of 17 Plasmodium species encompassing a wide genomic GC content range. Genic GC content was sorted and divided into ten equally sized quantiles that were then assessed for functional enrichment classes. In agreement that selection on gene classes may drive genomic GC content, transmembrane proteins were enriched within extreme GC content quantiles (Q1 and Q10). Specifically, variant surface antigens, which primarily interact with vertebrate immune systems, showed skewed GC content distributions compared with other trans-membrane proteins. Although a definitive causation linking GC content, expression, and positive selection within variant surface antigens from Plasmodium vivax, Plasmodium berghei, and Plasmodium falciparum could not be established, we found that regardless of genomic nucleotide composition, genic GC content and expression were positively correlated during trophozoite stages. Overall, these data suggest that, alongside mutational biases, functional protein classes drive Plasmodium GC content change.

Key words: GC content, GC bias, Plasmodium, CoGe, comparative genomics, gene function, genome evolution.

Introduction

The genus Plasmodium is formed of vector-borne obligate parasites capable of infecting a wide range of vertebrates. Notably, five species—Plasmodium vivax, Plasmodium knowlesi, Plasmodium malariae, Plasmodium ovale, and Plasmodium falciparum—cause disease in humans (malaria). Despite control and eradication efforts, malaria continues to have an enormous impact on global health. In 2015 alone, the World Health Organization reported an estimated 1 million new malaria cases and over 400,000 malaria-related deaths (WHO 2016). Furthermore, Plasmodium parasites negatively affect endemic biodiversity. Specifically, the introduction of Plasmodium relictum to the Hawaiian islands has caused significant declines in native bird population size (Atkinson et al. 2010; Samuel et al. 2015, 2017). In conjunction with their significance as human and animal pathogens, Plasmodium parasites are also valuable models for studying fundamental aspects of genome sequence evolution.

Perhaps one of the most noteworthy features of Plasmodium genomes is their rapid change in genomic Guanine–Cytosine (GC) content. Within the span of ~35–40 Myr (Pacheco et al. 2011), GC content has increased by roughly 30% across the genus. For example, ancestral species such as Plasmodium gabori have a genomic GC content of 17.85%, whereas more recently divergent species such as P. vivax have a genomic GC content of 44.87% (Carlton et al. 2008;
Sundararaman et al. 2016). Interestingly, coding sequence has a higher GC content than noncoding sequence in all Plasmodium genomes thus far sequenced (Castillo et al. 2018).

It has been suggested that mutational biases are driving this genome-wide nucleotide composition change. In a study using mutation accumulation laboratory lines, researchers proposed that an overrepresentation of spontaneous A/T substitutions has pushed the P. falciparum genome sequence toward higher AT content (Hamilton et al. 2016). Species closely related to P. falciparum (Laverania subgenus) are equally AT rich, implying that the substitution bias predates the divergence of the Laverania group (Hamilton et al. 2016). It has been hypothesized that a later shift in this mutational bias occurred sometime during Plasmodium evolution. Explicitly, P. vivax is thought to be subject to increased GC-biased heteroduplex repair and decreased AT-biased mutation (Nikbakht et al. 2015); however, the exact evolutionary mechanisms driving this shift are unknown. Although mutational biases have a clear role in shifting Plasmodium GC content, several unanswered questions remain. The evolutionary origin, causes, and mechanisms behind these mutational biases remain poorly understood. Furthermore, we lack insight in how the extreme changes in nucleotide composition influence transcription, translation, and protein structure across orthologs, and how different gene functions might favor or be resilient to nucleotide changes.

In other biological systems, natural selection is thought to have partly driven changes in GC content within certain protein families. For instance, in bacterial genomes, the core genome (genes present in all evaluated strains) is significantly more GC rich than the accessory genome (genes absent in certain strains), largely due to purifying selection constraining GC content on core genes (Bohlin et al. 2017). In humans, functional classes such as “Transcription,” “Signaling and transduction,” “Carbohydrate transport and metabolism,” “Amino acid transport and metabolism,” and “Inorganic ion transport and metabolism” have higher average GC content at the third nucleotide position (GC3) than the rest of the genome (D’Onofrio et al. 2007). Other studies have also suggested that lineage-specific differences in genomic GC content might be adaptive to environmental changes (Luo et al. 2015). For example, in plant monocots, it is believed that GC content may have shifted in response to desiccation stress in cold and dry climates (Smarda et al. 2014). These studies clearly show that in certain lineages and gene functional classes, changes in GC content can be adaptive.

Compared with other eukaryotes, the shift in nucleotide composition seen in the genus Plasmodium is unparalleled. Previous evidence shows that Plasmodium parasites have developed adaptive strategies to successfully infect members of their complex life cycle: the Anopheles vector and the vertebrate hosts (Molina-Cruz and Barillas-Mury 2014; Scully et al. 2017). Moreover, frequent vertebrate host-switches during Plasmodium evolution have also created an opportunity for adaptive change to occur (Stoltzfus and McCandlish 2017). Whether such adaptive change could influence the genome-wide compositional change seen in the genus Plasmodium remains to be fully explored.

Within the Plasmodium genus, certain protein families are strongly affected by natural selection. For example, Plasmodium antigen evolution is thought to occur under intense selective pressure imposed by host immunity (Loy et al. 2017). Among the proteins interacting with the vertebrate immune systems, variant surface antigens (VSAs) have a critical role in parasite virulence and immune evasion. VSAs are gene families characterized by multiple exons, accelerated sequence evolution, rapid gene turnover, and expression encompassing the trophozoite through schizont stages of the Plasmodium intraerythrocytic developmental cycle (Chen et al. 1998; Bozdech et al. 2008; Frech and Chen 2013; Yam et al. 2016; Martins et al. 2017). Even though they might share a common evolutionary origin (Janssen et al. 2002, 2004), VSA family members are largely species specific and share little sequence similarity between species. VSAs include species-specific families such as the Plasmodium Interspersed Repeats (pir) family: P. vivax (vir) and Plasmodium berghei (bir); the P. falciparum-specific Repetitive Interspersed Family (rif) coding for the rifin protein; the Subtelomeric Variant Open Reading frame (stevor) family; and var genes encoding the Erythrocyte Membrane Protein 1 (PfEMP1) family. Thus, the VSA proteins may be ideal candidates for driving genome sequence evolution in the genus Plasmodium.

Here, we evaluate the hypothesis that the rapid shift in genomic GC content seen in Plasmodium parasites is driven, in part, by adaptive changes in certain functional classes of genes. We characterized the contribution of genes belonging to different functional groups to the overall variation of coding GC content seen across Plasmodium lineages. Among all assessed groups, we found that trans-membrane proteins, particularly VSAs, had extreme GC content values compared with other functional classes within a species. Trans-membrane proteins represent ~20–30% of protein coding genes in most Plasmodium species (e.g., P. vivax [18.89%], P. berghei [38.60%], and P. falciparum [22.78%]) and contribute ~10–15% of the bases in the Plasmodium genomes (e.g., P. vivax [12.24%], P. berghei [10.16%], and P. falciparum [14.57%]). Consequently, we examined the putative relation between VSA expression during the trophozoite stage of the Plasmodium life cycle, signatures of selection, and genic GC content variations.

Materials and Methods

Mining Plasmodium’s Genomic, Coding, GC3, and Genic GC Content Values with CoGe

Representative genomes from the four major Plasmodium clades (simian, rodents, Laverania subgenus, and birds/ reptiles) were obtained from NCBI/GenBank (Clark et al. 2016), PlasmoDB (Aurrecoechea et al. 2009), and GeneDB.
Clustering defined by the “Trans-membrane/Integral component of membrane” clusters members of the VSA multigene families showed the most extreme GC content values. For each species, the differences in GC content distribution between VSA families and an equally sized randomized sample taken from the rest of the coding genome were evaluated with a Wilcoxon rank sum test in RStudio v3.3 (supplementary fig. S7, Supplementary Material online).

GC Content Distribution and Natural Selection on VSA Families

We assessed if there were signs of positive selective pressure that could hypothetically influence GC content within these families. Species-specific multiple sequence alignments were performed on the following VSA families: vir (P. vivax Salvador-1 strain [Carlton et al. 2008]), bir (P. berghei ANKA strain [Khasawneh and Kornreich 2005]), and var (P. falciparum 3D7 strain [Gardner et al. 2002]).

Codon alignments for the complete CDS of VSA sequences were performed using the MUSCLE algorithm (Edgar 2004) found on MEGA v7 (Kumar et al. 2016). Pseudogenes and incomplete CDS (e.g., sequences missing an exon) were excluded. Maximum likelihood (ML) trees were generated for each VSA family alignment using PhyML (Guindon et al. 2010). Each ML tree used a general time reversible model and 500 bootstrap replicates to evaluate node confidence (fig. 4a-c and supplementary file S8, Supplementary Material online). A Random Effects Branch-Site model (aBSREL) (Kosakovskyy Pond et al. 2011) was employed to detect signatures of positive selection in all branches. The analysis was run using Hyphy’s executable for the command line (Kosakovskyy Pond et al. 2005).

GC Content Distribution and Gene Expression during the Trophozoite Life Stage

The relationship between gene expression and GC content was evaluated in the three Plasmodium species with publicly available RNA-Seq data: P. vivax strain Salvador-1 (Zhu et al. 2016), P. berghei strain ANKA (Otto et al. 2014), and P. falciparum strain 3D7 (Otto et al. 2010). A significant portion of the Plasmodium life cycle is spent consecutively infecting and replicating within vertebrate’s erythrocytes (intraerythrocytic developmental cycle). During this phase, Plasmodium parasites are metabolically dependent on the infected erythrocyte...
FIG. 1.—Characterization of GC content in 17 Plasmodium genomes. (a) Plasmodium phylogeny: simian clade (green), rodent clade (red), Laverania subgenus (blue), and Plasmodium relictum from the birds/reptiles clade (yellow). (b) Histogram of genomic, coding, and GC3 content across 17 Plasmodium species: genomic GC (purple), coding GC (orange), and GC3 (teal). (c) GC content distribution of Plasmodium CDS. Simian clade: green, rodent clade: red, Laverania: blue, and P. relictum: yellow.

FIG. 2.—Histograms of GO-term enrichment within GC content quantiles. Histogram of genes (y axis) for bins of GC content (x axis). Vertical lines in each histogram delineate GC content quantiles. The number of genes associated with 11 GO-terms is shown as stacked colored bars. (a) Plasmodium vivax from the simian clade. (b) Plasmodium berghei from the rodent clade. (c) Plasmodium falciparum from the Laverania subgenus.
and are exposed to host immunity. VSA expression peaks in the Plasmodium intraerythrocytic developmental cycle during the trophozoite stage (Chen et al. 1998; Bozdech et al. 2008; Martins et al. 2017). Thus, the Fragments Per Kilobase of transcript per Million mapped reads (FPKM) values for the trophozoite stage were downloaded from PlasmoDB and log 10 transformed. A scatterplot of the genic GC content and log 10 FPKM values was created in RStudio v3.3, and the relation between these two variables was evaluated with a Pearson correlation test. This analysis was conducted both for VSA family members (fig. 5) and for all genes expressed during the trophozoite stage (supplementary fig. S9, Supplementary Material online).

**Results**

**Changes in Plasmodium Nucleotide Composition Accelerate in the Simian Clade**

Coding and GC content at the third nucleotide position (GC3) increased by 25% ± 4.26 and 40% ± 6.87, respectively, between ancestral (e.g., P. relictum) and recently diverged (e.g., P. vivax) species (fig. 1a and b). With the exception of P. malariae (Rutledge et al. 2017), GC content increased between 10 and 14% ± 3.76 in the simian clade compared with other Plasmodium clades. These variations were less pronounced (~1–3% ± 0.44–0.67) within and between the rodent clade, Laverania subgenus, and P. relictum (fig. 1b). Intraclade variations of genomic, coding, and GC3 content were noticeable in the simian clade (>10% ± 3.76), even when P. malariae was treated as an outlier (>10% ± 2.83) (fig. 1b). Moreover, the distribution of GC3 content was roughly bimodal in three simian clade species: P. vivax, Plasmodium cynomolgi (Tachibana et al. 2012), and Plasmodium coatneyi (Kissinger et al. 2016) (supplementary fig. S10, Supplementary Material online). Overall, both genome-wide coding and GC3 composition are rapidly shifting in the genus, particularly within the simian clade.

Distinct Functional Classes Uniquely Contribute to Changes in Plasmodium GC Content

For all species, the coding GC content displayed a nonstandard distribution that trended toward lower GC values for GC-poor genomes and higher GC values for GC-rich genomes (fig. 1c and supplementary table S2, Supplementary Material online). In addition, in some species, the coding GC content distribution was much broader than in others (supplementary table S2, Supplementary Material online). For example, GC-rich species P. vivax (46.7 ± 0.19), P. cynomolgi (42.63 ± 0.18), and Plasmodium fragile (42.11 ± 0.16) had broad distributions relative to other GC-rich species such as P. knowlesi (40.23 ± 0.11) (Pain et al. 2008) and P. coatneyi (41.95 ± 0.11) (fig. 1c). In sum, coding GC content is rapidly changing in the genus, even between closely related Plasmodium species of the simian clade.

The contribution of particular functional classes of genes to the observed shifts in GC content was assessed (fig. 2). The number of genes in Plasmodium species ranges between 5,245 and 5,992 (P. berghei and Plasmodium reichenowi, respectively (Sundararaman et al. 2016)). For each species, the CDS were divided into ten quantiles based on GC content. Quantile 1 (Q1) and Q10 represented the ~500–600 genes with the most extreme GC content (low or high, respectively) in each genome sequence annotation. The genes within each
quantile were analyzed for enrichment in specific functional classes (fig. 2a–c). Specific GO-terms were enriched in quantiles representing the most extreme GC content values.

In three species of the simian clade (P. vivax [fig. 2a], P. cynomolgi [supplementary fig. S7, Supplementary Material online], and Plasmodium inui [supplementary tables S3 and S5, Supplementary Material online]), GO-terms associated with “Trans-membrane/Integral component of membrane” terms were enriched in quantile 1 (Q1). Specifically, the ESs for the “Trans-membrane/Integral component of membrane” term in these species were P. vivax (ES = 28.4, P value = 7.1 e-32), P. cynomolgi (ES = 7.82, P value = 1.4 e-09), and P. inui (ES = 6.44, P value = 5.2 e-09). These genes represent 1.82%, 1.05%, and 1.05% of the total DNA content in the P. vivax, P. cynomolgi, and P. inui genomes, respectively. Q1 represents the set of genes with the lowest GC content for each species. Alternatively, in P. falciparum (fig. 2c), the closely related term “VSA/Trans-membrane/Integral component of membrane” was enriched in both Q1 (ES = 4.17, P value = 8.1 e-07; 2.27% of the genome) and Q10 (ES = 7.59, P value = 1.9 e-16; 0.76% of the genome), the latter represents the set of genes with the highest GC content for each species. This result is noteworthy because other enriched GO-terms, that is, “Translation and structural constituents of the ribosome” (P. berghei [Q8’s ES = 3.83, P value = 1.6 e-09; Q9’s ES = 3.89, P value = 3.3 e-08; Q10’s ES = 7.6, P values = 8.7 e-16]) and “DNA-binding and nucleosome core formation” (P. berghei [Q9’s ES = 6.19, P value = 4 e-13; Q10’s ES = 4.91, P values = 2.9 e-12]; P. falciparum [Q10’s ES = 3.5, P values = 7 e-05]) did not follow any similar pattern among the examined Plasmodium species (supplementary tables S3–S5, Supplementary Material online).

Given their contribution to extreme GC content values in many Plasmodium species, genes with an associated “Trans-membrane/Integral component of membrane” term were further clustered based on described gene function (fig. 3). For all evaluated Plasmodium species, the largest clusters were composed of trans-membrane proteins with unknown or
hypothetical functions (fig. 3a–c). Among all gene clusters with identified functions (supplementary table S6, Supplementary Material online), only the VSA group appears to show skewed GC content distributions compared with the rest of “Trans-membrane/Integral component of membrane” genes (fig. 3a–c). *Plasmodium vivax*’s VSA family members (vir) had lower GC content values compared with other “Trans-membrane/Integral component of membrane” associated genes ($t = -29.83, P\text{ value} = 2.2 \times 10^{-16}$, fig. 3a). In contrast, *P. falciparum* VSA families (var, stevor, and rifin) had higher GC content values in comparison to the rest of “Trans-membrane/Integral component of membrane” associated genes ($t = 40.31, P\text{ value} = 2.2 \times 10^{-16}$, fig. 3c). A similar trend was observed for VSA family members in taxa related to *P. vivax* and *P. falciparum*, respectively. Specifically, VSA family members found in *P. cynomolgi* (sister taxa to *P. vivax*) also had lower coding GC values compared with the rest of the genome sequence (supplementary fig. S7b, Supplementary Material online). In contrast, VSA family members found in *Plasmodium reichenowi* and *Plasmodium gaboni* (closely related to *P. falciparum*) showed higher coding GC compared with the rest of the genome sequence (supplementary fig. S7b, Supplementary Material online). A Wilcoxon rank sum test showed that the compositional differences between VSA family members and the rest of the coding genome were significant for most species (supplementary fig. S7a, Supplementary Material online).

**Gene Expression and VSA GC Content, but Not Positive Selection, Are Positively Correlated**

Though GC content is shifting in *Plasmodium* parasites, previous studies indicate that a significant portion of the *Plasmodium* genome evolves under strong constraints and
purifying selection (Cornejo et al. 2014). Moreover, purifying selection is thought to be intensified by the complex *Plasmodium* parasitic life cycle (Chang et al. 2013). Contrary to the rest of the genome, positive selection has been detected in genes related to the development of antimalarial drug resistance, parasite development in mosquitoes, and variant-expressed multigene families (Park et al. 2012; Cornejo et al. 2014; Duffy et al. 2015). Among these, VSA families are thought to evolve under strong host imposed immune-response selective pressures (Loy et al. 2017); thus, putative signatures of positive selection within VSA family trees were evaluated. In addition, because variations in GC content are known to affect gene expression (Newman et al. 2016), the expression for these gene families was also examined using publicly available RNA-Seq data for the trophozoite stage of *P. vivax*, *P. berghei*, and *P. falciparum* (Otto et al. 2010, 2014; Zhu et al. 2016).

The correlation between VSA genes under positive selection and their expression was assessed (figs. 4 and 5). Expressed genes and phylogenetic branches showing significant signs of positive selection were interspersed across both phylogeny and GC content values (fig. 4a–c). Significant signs of positive selection were found in the var family tree (fig. 4c) in accordance with previous reports (Plotkin et al. 2004). However, this family also presents the largest number of identified recombination breakpoints compared with other VSAs. It has been suggested that both forces act simultaneously in var genes, with mitotic recombination facilitating gene diversity, and positive selection favoring gene variants better equipped to evade host’s immune responses (Bopp et al. 2013; Claessens et al. 2014). Genic GC content was significantly higher for VSA *P. falciparum* genes showing signs of positive selection compared with nonselected genes (t = 9.893, P value = 6.76 e-16). It was not possible to compare this trend in *P. berghei* or *P. vivax* due to the small number of sequences with signatures of positive selection detected. VSA GC content and expression during the trophozoite stage was positively correlated in both *P. falciparum* (r = 0.282, P value = 1.06 e-5) and *P. vivax* (r = 0.364, P value = 3.36 e-8), but not in *P. berghei* (fig. 5). Similarly, a significant and positive correlation between GC content and expression during the trophozoite life cycle stage was observed when examining all genes in all three species (supplementary fig. S9a–c, Supplementary Material online).

**Discussion**

Here, a large-scale comparative genomic approach was used to examine factors influencing observed shifts in genomic GC content within the *Plasmodium* genus (fig. 1 and supplementary table S7, Supplementary Material online). In agreement with previous reports (Castillo et al. 2018), coding GC content in all *Plasmodium* spp. was found to be elevated in comparison to genomic GC content. An interesting pattern where GC3 content was higher than genomic and coding GC content in the simian clade, *Laverania* subgenus, and the ancestral lineage, *P. relictum*, was also observed. In eukaryotes, protein expression (Yadav and Swati 2012), function (Louie et al. 2003), and structure (Basile et al. 2008) have been known to limit coding GC content, partly due to selective constraints. These selective constraints may result from mutational biases, such as the elevated rate of G/C to A/T transitions compared with the rate of A/T transitions to G/C observed in *P. falciparum* mutation accumulation lines (Hamilton et al. 2016) or the elevated GC3 content in *P. vivax* (Nikbakht et al. 2015). The shifts in GC3 content observed here suggest that putative differences in mutational bias are not limited to either *P. vivax* or *P. falciparum*, but instead show a strong phylogenetic pattern and likely originate from ancestral events during the evolution of the genus. Furthermore, the bimodal GC3 distributions observed in some *Plasmodium* species (supplementary fig. S10, Supplementary Material online) hint to groups of genes with distinct mutational biases or selective pressures within a single species.

Genes with associated GO-terms, “Trans-membrane/Integral component of membrane” was the dominant GO-term recovered for most GC content quantiles (fig. 2). Moreover, this GO-term, which contains genes essential for parasite survival (Sullivan and Krishna 2005; Curtidor et al. 2008; Kirk and Lehane 2014), was highly enriched in boundary GC content quantiles (Q1 and Q10), especially in species with extreme nucleotide compositions. During erythrocyte invasion and intraerythrocytic development (Curtidor et al. 2008), some *Plasmodium* integral membrane proteins are exposed to the immune system, thus providing an opportunity for strong positive selection. In addition, proteins with trans-membrane motifs evolve significantly faster than other *Plasmodium* genes (Carlton et al. 2008). A further examination of this GO-term class revealed that, with some exceptions, GC content was particularly skewed to the extremes in the VSA family of trans-membrane proteins, though a significant number of trans-membrane proteins are annotated as “unknown/hypothetical proteins” (fig. 3). Similar to our observation in species of the subgenus *Laverania*, in *Haemoproteus tartakovskyi* (a bird-infective obligate parasite closely related to the genus *Plasmodium*), VSA family members are enriched for higher GC content relative to other coding genes (Bensch et al. 2016). In contrast, VSA family members of the simian clade harbored a lower GC content relative to the rest of the genome sequence (fig. 3). Thus, regardless of the species, it is clear that VSA genes display GC contents that seem to be at odds with the rest of the genome sequence. This suggests that host-related selective pressures drive rapid sequence evolution, which may drive genomic GC content change in those genes. Nonetheless, it remains unclear if the selection acting on VSA is driving species-level shifts in genomic GC content, or if GC content on these genes is simply lagging behind the rest of the genome sequence.
Attempts to address this question were confounded by
the lack of a clear relationship between VSA GC content
and signatures of selection at the genic level (fig. 4a–c).
Significant signs of positive selection were interspersed in
both internal and terminal branches of each phylogeny,
regardless of genic GC content at the tips. This suggests
that the enrichment of VSA genes in extreme GC quanti-
ties is not a result of selection along specific phylogenetic
branches of gene paralogs. The evidence of selection
revealed is likely linked to specific gene functions instead
and may hint at rapid diversifying selection operating on
many genes simultaneously. An alternative explanation
for why VSA GC content appears to be evolving in the
opposite direction of other genes may be found in non-
adaptive forces such as recombination, GC-biased gene
conversion, and local nucleotide biases. Both mitotic re-
combination (Bopp et al. 2013; Claessens et al. 2014) and
variation in local nucleotide composition (Carlton et al.
2008; Tachibana et al. 2012) have been considered as
potential factors mediating VSA evolution and might re-
sult in the trends observed here.

Interestingly, with the exception of P. knowlesi (Pain et al.
2008), P. coatneyi (Kissing et al. 2016), and species from
the subgenus Laverania, VSA families are located in subtelomeric
chromosome regions (Gardner et al. 2002; Carlton et al.
2008; Tachibana et al. 2012; Rutledge et al. 2017). Nucleotide
composition in these regions is strongly AT biased
in P. vivax and P. cynomolgi (Carlton et al. 2008; Tachibana
et al. 2012), but not so in other Plasmodium species (supple-
mentary fig. S11, Supplementary Material online). Significant
changes in VSA nucleotide composition compared with the
rest of the coding genome were observed in Plasmodium
species with VSAs clustered in the subtelomeric chromosome
regions as well as in species with VSAs distributed across
the chromosome length (e.g., P. falciparum and P. knowlesi). In
species with significantly GC-poor subtelomeres (P. vivax and
P. cynomolgi), VSAs were the most numerous protein group
within 30 kb of the chromosome ends. In addition, the GC
content distribution within this region emulated that previ-
ously observed for the Trans-membrane/Integral component
of membrane groups, specifically VSAs (figs. 2a and 3a). On
the other hand, in GC-rich Plasmodium species with clustered
subtelomeric VSAs (e.g., P. malariæ) and species with scat-
tered VSAs (e.g., P. knowlesi and P. coatneyi), the GC content
distribution of CDS within the same 30 kb region emulated
that observed on the entire coding genome (supplementary
fig. S12, Supplementary Material online). Nonetheless VSAs in
these species had significantly smaller GC content differences
compared with the rest of the genome (supplementary fig.
S7, Supplementary Material online). Overall, though there are
some species-specific variations, the GC content of VSA fam-
ily members remains at odds with the GC content of the rest
of the genome sequence, irrespective of their genomic
location.

Whether the lower GC content observed in P. vivax and P.
cynomolgi VSA families originates from unique AT-biased mu-
tation patterns within these regions, or if VSA-specific pro-
cesses drive subtelomeric GC content to lower values is
something that remain inconclusive. Regions of lower amino
acid diversity compared with other coding regions have pre-
viously been described in P. falciparum (Zilversmit et al. 2010).
Moreover, they have been commonly described in intragenic
regions of VSA families across Plasmodium species (Janssen
et al. 2004; Joannin et al. 2008). Recent studies have found
that regions of low complexity tend to remain unchanged
over evolutionary time and that composition preferences in
the region are unrelated to compositional biases observed in
the rest of the genome (Chaudhry et al. 2018). Furthermore,
the authors suggest that natural selection might play a key
role in determining the amino acid composition in regions of
low complexity (Chaudhry et al. 2018). Therefore, it is possible
that the nucleotide composition differences in VSAs com-
pared with the rest of the coding genome might be the prod-
uct of selective forces acting on low complexity region across
Plasmodium species. Previous work showed that highly
expressed genes favor usage of AT-ending codons in both
P. falciparum and P. vivax (Yadav and Swati 2012). Thus,
the positive correlation between VSA expression and their
GC content observed in P. vivax and P. falciparum during the
trophozoite stage of the Plasmodium life cycle is noteworthy
(fig. 5) because it contradicts the general trend observed
across all genes. P. vivax and Plasmodium chabaudi
(Khasawneh and Komreik 2005), some VSAs mediate
cytoadherence (Bernabeu et al. 2012), rosette formation,
and cell invasion (Yam et al. 2016) and are trafficked to dif-
ferent cellular compartments. A preference for higher GC
content in certain VSA family members may thus reflect dif-
ferential subcellular erythrocyte localization and their unique
functions (Bernabeu et al. 2012). In addition, intrafamily dif-
fences in GC content might be connected to a proposed
strategy whereby P. falciparum alters expression patterns of
VSA genes to avoid immunological detection and promote
survival (Abdi et al. 2016).

Previous studies have led to contradictory conclusions
regarding the genome-wide relation between GC content
and gene expression. Although some studies show a neg-
ative correlation between coding GC content and gene
expression in eukaryotes (Rao et al. 2013), others have
described a positive correlation between both factors
(Kudla et al. 2006). Despite the drastic differences in GC
content observed within the Plasmodium genus, all tested
species displayed a preference for high GC content during
the trophozoite stage, which aligns with the latter model
(supplementary fig. S9, Supplementary Material online).
Previous studies have proposed that Plasmodium transcrip-
tion is highly optimized and that most genes function at
multiple points in the life cycle (Bushell et al. 2017).
Therefore, it is possible that Plasmodium genic GC content

505

Genome Biol. Evol. 11(2):497–507 doi:10.1093/gbe/evz015 Advance Access publication January 23, 2019

Tail Wags the Dog?
changes could be related to mRNA transcription and processing efficiency. Alternatively, although other studies suggest that translational selection has a limited role in shaping global codon and amino acid usage bias within the genus *Plasmodium*; the correlation between amino acid frequency of highly expressed genes and their respective tRNA is slightly higher in some species (Yadav and Swati, 2012). Therefore, it would be of interest to further evaluate this pattern specifically in the VSA gene family.

Given the importance of efficiently expressing VSA genes, their translational efficiency may be driving genome-wide changes in GC content. Indeed, the evidence of biased GC content in the subtelomeric genomic regions harboring VSA genes (e.g., *P. vivax*) versus the rest of the genome may be a footprint of this effect. Although the presented data do not refute that hypothesis, additional population-level genomic and functional surveys as well as long-term evolution studies are needed in *Plasmodium*. Ideally, an artificial selection scheme similar to the long-term adaptive evolution studies in *Escherichia coli* (Blount et al. 2012) would be performed to definitively determine if the tail wags the dog for genomic GC composition evolution in *Plasmodium* species.

In sum, large-scale comparative genomic analyses are powerful means to understand how genome sequences evolve, particularly regarding genome-wide GC content shifts. Nucleotide compositional changes are crucial in genome evolution, phylogenetic reconstruction, and in evaluating expression and translation dynamics. *Plasmodium* genomes are incredibly valuable models to explore the causes and effects of such nucleotide composition shifts. Moreover, the availability of many sequenced genomes within the genus provides a unique resource to better understand the forces driving rapid shifts in genomic GC content. When attempting to understand what is driving differences in genomic GC content within the *Plasmodium* genus, these data suggest that, alongside mutational biases, functional protein classes drive GC content changes across the genome.

**Supplementary Material**

Supplementary data are available at Genome Biology and Evolution online.

**Acknowledgments**

This work was supported by the National Science Foundation awards: NSF IOS-1339156 and IOS-1444490. We thank CyVerse (NSF DBI-0735191 and DBI-1265383) for sharing best practices and providing data and computational infrastructure for CoGe (NSF IOS-1339156 and IOS-1444490). We are also grateful to Mark Beilstein and Asher K. Haug-Baltzell for their helpful comments and Sean Davey for providing assistance with CoGe’s resources.

**Literature Cited**

Abdi AI, et al. 2016. Global selection of *Plasmodium falciparum* virulence antigen expression by host antibodies. Sci Rep. 6:1–9.

Atkinson CT, Samuel MD, Atkinson CT, Samuel MD. 2010. Nordic Society Oikos Avian malaria *Plasmodium relictum* in native Hawaiian forest birds: epizootiology and demographic impacts on ‘apapane *Himatione Sanguinea*. J Avian Biol. 41:357–366.

Aurrecoechea C, et al. 2009. PlasmoDB: a functional genomic database for malaria parasites. Nucleic Acids Res. 37:539–543.

Basile W, Sachenkova O, Light S, Eloffson A. 2017. High GC content causes orphan proteins to be intrinsically disordered. PLoS Comput Biol. 13:1–19.

Bensch S, et al. 2016. The genome of *Haemoproteus tartakowskyi* and its relationship to human malaria parasites. Genome Biol Evol. 8:1361–1373.

Bernabeu M, et al. 2012. Functional analysis of *Plasmodium vivax* VIR proteins reveals different subcellular localizations and cytoadherence to the ICAM-1 endothelial receptor. Cell Microbiol. 14(3):386–400.

Blount ZD, Barrick JE, Davidson CJ, Lenski RE. 2012. Genomic analysis of a key innovation in an experimental *Escherichia coli* population. Nature 489(7417):513.

Bohlin J, Eldholm V, Pettersson JHO, Brynildsrud O, Snipen L. 2017. The nucleotide composition of microbial genomes indicates differential patterns of selection on core and accessory genomes. BMC Genomics 18:151.

Bopp SE, et al. 2013. Mitotic evolution of *Plasmodium falciparum* shows a stable core genome but recombination in antigen families. PLoS Genet. 9(2):e1003293.

Bozdech Z, et al. 2008. The transcriptome of *Plasmodium vivax* reveals divergence and diversity of transcriptional regulation in malaria parasites. Proc Natl Acad Sci U S A. 105(42):16290–16295.

Bushell E, et al. 2017. Functional profiling of a *Plasmodium* genome reveals an abundance of essential genes. Cell 170(2):260–272.e18.

Carlton JM, et al. 2008. Comparative genomics of the neglected human malaria parasite *Plasmodium vivax*. Nature 455(7214):757–763.

Castillo AI, Nelson ADL, Haug-Baltzell AK, Lyons E. 2018. A tutorial of diverse genome analysis tools found in the CoGe web-platform using *Plasmodium* spp. as a model. Database 2018:1–16.

Chang H-H, et al. 2013. Malaria life cycle intensifies both natural selection and random genetic drift. Proc Natl Acad Sci U S A. 110(50):20129–20134.

Chaudhry SR, Lwin N, Phelan D, Escalante AA, Battistuzzi FU. 2018. Comparative analysis of low complexity regions in *Plasmodia*. Sci Rep. 8(1):335.

Chen Q, et al. 1998. Developmental selection of var gene expression in *Plasmodium falciparum*. Nature 394(6691):392–395.

Claessens A, et al. 2014. Generation of antigenic diversity in *Plasmodium falciparum* by structured rearrangement of Var genes during mitosis. PLoS Genet. 10(12):e1004812.

Clark K, Karsch-Mizrachi I, Lipman DJ, Ostell J, Sayers EW. 2016. GenBank. Nucleic Acids Res. 44(D1):D67–D72.

Cornejo OE, Fisher D, Escalante AA. 2014. Genome-wide patterns of genetic polymorphism and signatures of selection in *plasmodium vivax*. Genome Biol Evol. 7:106–119.

Curtidor H, et al. 2008. Characterization of *Plasmodium falciparum* integral membrane protein Pf25-IMP and identification of its red blood cell binding sequences inhibiting merozoite invasion in vitro. Protein Sci. 17(9):1494–1504.

D’Onofrio G, Ghosh TC, Saccone S. 2007. Different functional classes of genes are characterized by different compositional properties. FEBS Lett. 581:5819–5824.

Duffy CW, et al. 2015. Comparison of genomic signatures of selection on *Plasmodium falciparum* between different regions of a country with high malaria endemicity. BMC Genomics. 16:1–11.
Tail Wags the Dog?

Edgar RC. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. 32(5):1792–1797.

Freh C, Chen N. 2013. Variant surface antigens of malaria parasites: functional and evolutionary insights from comparative gene family classification and analysis. BMC Genomics. 14(1):427.

Gardner MJ, et al. 2002. Genome sequence of the human malaria parasite Plasmodium falciparum. Nature 419(6906):498–511.

Guindon S, et al. 2010. New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. Syst Biol. 59(3):307–321.

Hamilton WL, et al. 2016. Extreme mutation bias and high AT content in Plasmodium falciparum. Nucleic Acids Res. 45(4):1889–1901.

Haug-Baltzell AK, Davey S, Scheidegger C, Lyons E. 2017. SynMap2 and SynMap3D: web-based whole-genome synteny browsers. Bioinformatics 33:2197–2198.

Kissinger JC, et al. 2016. High-quality genome assembly and annotation of large gene lists using DAVID bioinformatics resources. Nat Protoc. 4(1):44–57.

Jansen CS, Barrett MP, Turner CM, Phillips RS. 2002. A large gene family for putative variant antigens shared by human and rodent malaria parasites. Proc Biol Sci. 269(1489):431–436.

Kudla G, Lipinski L, Caffin F, Helwak A, Zylicz M. 2006. High guanine and cytosine content increases mRNA levels in mammalian cells. PloS Biol. 4(9):383–394.

Kumar S, Stecher G, Tamura K. 2016. MEGA7: molecular evolutionary genetics analysis version 7.0 for bigger datasets. Mol Biol Evol. 33(7):1870–1874.

Logan-Klumperl FI, et al. 2012. GeneDB—an annotation database for interspersed repeats: the major multigene superfamily of malaria parasites. Nucleic Acids Res. 32(19):5712–5720.

Otto TD, et al. 2014. A comprehensive evaluation of rodent malaria parasite genomes and gene expression. BMC Biol. 12:86.

Pacheco MA, et al. 2011. Timing the origin of human malarias: the lemur puzzle. BMC Evol Biol. 11(1):299.

Pain A, et al. 2008. The genome of the simian and human malaria parasite Plasmodium knowlesi. Nature 455(7214):799–803.

Park DJ, et al. 2012. Sequence-based association and selection scans identify drug resistance loci in the Plasmodium falciparum malaria parasite. Proc Natl Acad Sci U S A. 109(32):13052–13057.

Ploekin JB, Dushoff J, Fraser HB. 2004. Detecting selection using a single genome sequence of M. tuberculosis and P. falciparum. Nature 428(7086):942–945.

Rao YS, Chai XW, Wang ZF, Nie QH, Zhang XQ. 2013. Impact of GC content on gene expression pattern in chicken. Genet Sel Evol. 45:1–7.

Rutledge GG, et al. 2017. Plasmodium malariae and P. ovale genomes provide insights into malaria parasite evolution. Nature 542(7639):101–104.

Samuel MD, Woodworth BL, Atkinson CT, Hart PJ, LaPointe DA. 2015. Avian malaria in Hawaiian forest birds: infection and population impacts across species and elevations. Ecosphere 6:1–21.

Samuel MD, et al. 2017. The dynamics, transmission, and population impacts of avian malaria in native Hawaiian birds: a modeling approach. Ecol Appl. 21:2960–2973.

Scully EJ, Kanjee U, Duraasingh MT. 2017. Molecular interactions governing host-specificity of blood stage malaria parasites. Curr Opin Microbiol. 40:21–31.

Smarda P, et al. 2014. Ecological and evolutionary significance of genomic GC content diversity in monocots. Proc Natl Acad Sci U S A. 111(39):E4096–E4102.

Stoltzfus A, McCardlsh DM. 2017. Mutational biases influence parallel adaptation. Mol Biol Evol. 34(9):2163–2172.

Sullivan DJ, Krishna S. 2005. Malaria drugs, disease, and post-genomic biology. Berlin Heidelberg New York: Springer.

Sundararaman SA, et al. 2016. Genomes of cryptic chimpzæe Plasmodium species reveal key evolutionary events leading to human malaria. Nat Commun. 7:11078.

Tachibana S-I, et al. 2012. Plasmodium cynomolgi genome sequences provide insight into Plasmodium vivax and the monkey malaria clade. Nat Genet. 44(9):1051–1055.

Tang H, Lyons E. 2012. Unleashing the genome of Brassica Rapa. Front Plant Sci. 3:1–12.

WHO. 2016. World malaria report. Genève, Switzerland: Global Malaria Programme World Health Organization.

Yadav MK, Swati D. 2012. Comparative genome analysis of six malarial parasites using codon usage bias based tools. Bioinformation 8(24):1230–1239.

Yam XY, et al. 2016. Characterization of the Plasmodium Interspersed Repeats (PIR) proteins of Plasmodium chabaudi indicates functional diversity. Sci Rep. 6:1–13.

Zhu L, et al. 2016. New insights into the Plasmodium vivax transcriptome using RNA-Seq. Sci Rep. 6:20498.

Zilversmit MM, et al. 2010. Low-complexity regions in Plasmodium falciparum: missing links in the evolution of an extreme genome. Mol Biol Evol. 9:2198–2209.

Associate editor: Geoff McFadden