A New Method for Estimating Species Age Supports the Coexistence of Malaria Parasites and Their Mammalian Hosts

Joana C. Silva,*1,2 Amy Egan,†2 Cesar Arze,1 John L. Spouge,3 and David G. Harris4

1Institute for Genome Sciences, University of Maryland School of Medicine
2Department of Microbiology and Immunology, University of Maryland School of Medicine
3National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, MD
4Department of Applied Mathematics and Statistics, University of Maryland, College Park
†Present address: Noblis, Falls Church, VA
*Corresponding author: E-mail: jcsilva@som.umaryland.edu.
Associate editor: Koichiro Tamura

Abstract

Species in the genus Plasmodium cause malaria in humans and infect a variety of mammals and other vertebrates. Currently, estimated ages for several mammalian Plasmodium parasites differ by as much as one order of magnitude, an inaccuracy that frustrates reliable estimation of evolutionary rates of disease-related traits. We developed a novel statistical approach to dating the relative age of evolutionary lineages, based on Total Least Squares regression. We validated our approach accurately reconstructs the age of well-established Drosophila clades, including the speciation event that led to the subgenera Drosophila and Sophophora, and age of the melanogaster species subgroup. We applied this approach to hundreds of loci from seven mammalian Plasmodium species. We demonstrate the existence of a molecular clock specific to individual Plasmodium proteins, and estimate the relative age of mammalian-infecting Plasmodium. These analyses indicate that: 1) the split between the human parasite Plasmodium vivax and P. knowlesi, from Old World monkeys, occurred 6.1 times earlier than that between P. falciparum and P. reichenowi, parasites of humans and chimpanzees, respectively; and 2) mammalian Plasmodium parasites originated 22 times earlier than the split between P. falciparum and P. reichenowi. Calibrating the absolute divergence times for Plasmodium with eukaryotic substitution rates, we show that the split between P. falciparum and P. reichenowi occurred 3.0–5.5 Ma, and that mammalian Plasmodium parasites originated over 64 Ma. Our results indicate that mammalian-infecting Plasmodium evolved contemporaneously with their hosts, with little evidence for parasite host-switching on an evolutionary scale, and provide a solid timeframe within which to place the evolution of new Plasmodium species.

Key words: Plasmodium, molecular clock, speciation dates, total least squares, regression, malaria, Drosophila.

Introduction

Malaria remains a leading cause of morbidity and mortality from infectious disease (Honey 2009), with over 200 million new cases and more than half a million deaths annually, despite increased efforts to control and eradicate the disease (World Health Organization 2009). The age of Plasmodium species informs our understanding of malaria transmission and, in particular, the likelihood of zoonosis. However, the timing of the divergence of these species from their close relatives remains highly controversial (Hayakawa et al. 2008; Rich et al. 2009; Hughes and Verra 2010; Ricklefs and Outlaw 2010; Tanabe et al. 2010; Silva et al. 2011). Estimates of the age of the split between Plasmodium falciparum and P. reichenowi (the latter a chimpanzee parasite) range from 5 to 7 Ma (Escalante et al. 1995; Hughes and Verra 2010; Silva et al. 2011), the estimated age of the split between their mammalian hosts (Steiper and Young 2006; Yang and Rannala 2006; Hobolth et al. 2007), to as recently as 10,000 years ago (Rich et al. 2009). The divergence between P. vivax and the Old World monkey parasite P. knowlesi has been estimated to date from 20 to 30 Ma (Escalante et al. 1995; Silva et al. 2011) to as recently as 2–3 Ma (Escalante et al. 1998). Likewise, the origin of the Plasmodium clade that parasitizes mammals, originally believed to date back ≥100 My (Escalante and Ayala 1995; Escalante et al. 1995), has also been placed within the last 13 My (Ricklefs and Outlaw 2010). The corresponding studies obtained their age estimates by converting genetic polymorphism or divergence into time, on the specific assumption of the coevolution of a host–parasite species pair, or of a substitution rate. They shared a major weakness in relying on a small number of loci, whose polymorphism and divergence might not be representative of the entire genome. The availability of complete or high quality (HQ) draft genomes from several mammalian malaria parasites overcomes this weakness. These species include the primate parasites P. falciparum, P. vivax, P. reichenowi, and P. knowlesi (Gardner et al. 2002; Jeffares et al. 2007; Carlton et al. 2008; Pain et al. 2008), and the three rodent parasites P. yoelli, P. chabaudi, and P. berghei (Carlton et al. 2002; Hall et al. 2005).
Here, our aim is to use genome-wide protein sequence divergence estimates to establish relative ages for specific speciation events occurring in the history of mammalian *Plasmodium*. Our basic premise is that if *Plasmodium* nuclear proteins evolve according to individual molecular clocks (Zuckerkandl and Pauling 1965; Kimura 1968), then sequence divergence in different proteins is correlated across independent *Plasmodium* lineages and, consequently, the regression slope of the divergence between the proteins in two lineages reflects the relative age of those lineages (fig. 1). We use *Plasmodium* genome sequences and the respective genome annotations to derive groups of orthologous single-copy genes across the seven species described above, and to obtain reliable estimates of divergence between protein sequences. In our statistical model, the clock for each protein has a specific rate, faster or slower, depending on the protein’s functional and structural constraints (Bromham and Penny 2003). To investigate the existence of a molecular clock in *Plasmodium*, the model derives regressions and $R^2$-statistics from the data, to quantify the relative rate of evolution of different proteins and determine whether the relative rates remain constant across lineages.

We demonstrate that the evolution of proteins encoded by single-copy genes in *Plasmodium* conforms remarkably well to a simple molecular clock model, permitting us to establish relative ages for speciation events among mammalian *Plasmodium* species accordingly. Finally, we convert relative ages to absolute divergence times using a range of well-to-a simple molecular clock, permitting us to establish relative ages for speciation events among mammalian *Plasmodium* species accordingly. Finally, we convert relative ages to absolute divergence times using a range of well-to-a simple molecular clock, permitting us to establish relative ages for speciation events among mammalian *Plasmodium* species accordingly. Finally, we convert relative ages to absolute divergence times using a range of well-to-a simple molecular clock, permitting us to establish relative ages for speciation events among mammalian *Plasmodium* species accordingly. Finally, we convert relative ages to absolute divergence times using a range of well-to-a simple molecular clock, permitting us to establish relative ages for speciation events among mammalian *Plasmodium* species accordingly. Finally, we convert relative ages to absolute divergence times using a range of well-to-a simple molecular clock, permitting us to establish relative ages for speciation events among mammalian *Plasmodium* species accordingly. Finally, we convert relative ages to absolute divergence times using a range of well-to-a simple molecular clock, permitting us to establish relative ages for speciation events among mammalian *Plasmodium* species accordingly. Finally, we convert relative ages to absolute divergence times using a range of well-to-a simple molecular clock, permitting us to establish relative ages for speciation events among mammalian *Plasmodium* species accordingly. Finally, we convert relative ages to absolute divergence times using a range of well-to-a simple molecular clock, permitting us to establish relative ages for speciation events among mammalian *Plasmodium* species accordingly. Finally, we convert relative ages to absolute divergence times using a range of well-to-a simple molecular clock, permitting us to establish relative ages for speciation events among mammalian *Plasmodium* species accordingly. Finally, we convert relative ages to absolute divergence times using a range of well-to-a simple molecular clock, permitting us to establish relative ages for speciation events among mammalian *Plasmodium* species accordingly. Finally, we convert relative ages to absolute divergence times using a range of well-to-a simple molecular clock, permitting us to establish relative ages for speciation events among mammalian *Plasmodium* species accordingly. Finally, we convert relative ages to absolute divergence times using a range of well-to-a simple molecular clock, permitting us to establish relative ages for speciation events among mammalian *Plasmodium* species accordingly. Finally, we convert relative ages to absolute divergence times using a range of well-to-a simple molecular clock, permitting us to establish relative ages for speciation events among mammalian *Plasmodium* species accordingly. Finally, we convert relative ages to absolute divergence times using a range of well-to-a simple molecular clock, permitting us to establish relative ages for speciation events among mammalian *Plasmodium* species accordingly. Finally, we convert relative ages to absolute divergence times using a range of well-to-a simple molecular clock, permitting us to establish relative ages for speciation events among mammalian *Plasmodium* species accordingly. Finally, we convert relative ages to absolute divergence times using a range of well-to-a simple molecular clock, permitting us to establish relative ages for speciation events among mammalian *Plasmodium* species accordingly. Finally, we convert relative ages to absolute divergence times using a range of well-to-a simple molecular clock, permitting us to establish relative ages for speciation events among mammalian *Plasmodium* species accordingly. Finally, we convert relative ages to absolute divergence times using a range of well-to-a simple molecular clock, permitting us to establish relative ages for speciation events among mammalian *Plasmodium* species accordingly. Finally, we convert relative ages to absolute divergence times using a range of well-to-a simple molecular clock, permitting us to establish relative ages for speciation events among mammalian *Plasmodium* species accordingly. Finally, we convert relative ages to absolute divergence times using a range of well-to-a simple molecular clock, permitting us to establish relative ages for speciation events among mammalian *Plasmodium* species accordingly. Finally, we convert relative ages to absolute divergence times using a range of well-to-a simple molecular clock, permitting us to establish relative ages for speciation events among mammalian *Plasmodium* species accordingly. Finally, we convert relative ages to absolute divergence times using a range of well-to-a simple molecular clock, permitting us to establish relative ages for speciation events among mammalian *Plasmodium* species accordingly. Finally, we convert relative ages to absolute divergence times using a range of well-to-a simple molecular clock, permitting us to establish relative ages for speciation events among mammalian *Plasmodium* species accordingly. Finally, we convert relative ages to absolute divergence times using a range of well-to-a simple molecular clock, permitting us to establish relative ages for speciation events among mammalian *Plasmodium* species accordingly. Finally, we convert relative ages to absolute divergence times using a range of well-to-a simple molecular clock, permitting us to establish relative ages for speciation events among mammalian *Plasmodium* species accordingly. Finally, we convert relative ages to absolute divergence times using a range of well-to-a simple molecular clock, permitting us to establish relative ages for speciation events among mammalian *Plasmodium* species accordingly. Finally, we convert relative ages to absolute divergence times using a range of well-to-a simple molecular clock, permitting us to establish relative ages for speciation events among mammalian *Plasmodium* species accordingly. Finally, we convert relative ages to absolute divergence times using a range of well-to-a simple molecular clock, permitting us to establish relative ages for speciat...
of magnitude, as would be expected for a diverse set of proteins exposed to a wide variety of selective constraints imposed by both structure and function. If *Plasmodium* proteins evolve according to a protein molecular clock, then the most conserved proteins in one lineage will also be conserved in other lineages and, conversely, rapidly evolving proteins will diverge rapidly in all lineages.

The Molecular Clock Model

To test the existence of protein-specific molecular clocks, we use the following model: let \( m_1, m_2, m_3 \) and \( m_4 \) be any four malarial species. Consider only species pairs \((m_1, m_2)\) and \((m_3, m_4)\) that lie on nonoverlapping tree-paths (i.e., separate branches) of the corresponding phylogenetic tree (fig. 1). The model assumes that each gene \( g \) common to the four malarial species has its own characteristic rate of amino acid substitution \( r(g) \), so that given the evolutionary time \((m_1, m_2)\) for divergence between the species \( m_1 \) and \( m_2 \), the amino acid sequence distance \( d(m_1, m_2; g) \) between the species \( m_1 \) and \( m_2 \) within the gene \( g \) (specified below) satisfies

\[
d(m_1, m_2; g) = t(m_1, m_2)r(g),
\]

that is, the amino acid sequence distance for a protein between two species is proportional to the species’ divergence time and the amino acid substitution rate of the protein. Because similar considerations apply to \( m_3 \) and \( m_4 \),

\[
\frac{d(m_3, m_4; g)}{d(m_1, m_2; g)} = \frac{t(m_3, m_4)r(g)}{t(m_1, m_2)r(g)} = \frac{t(m_3, m_4)}{t(m_1, m_2)}.
\]

Denote the final ratio by \( \alpha(m_1, m_2, m_3, m_4) \), a quantity reflecting the divergence time between \((m_3, m_4)\) in units of the divergence time between \((m_1, m_2)\). If our premise that a molecular clock exists is true, protein divergence in the two species pairs should be correlated, that is,

\[
d(m_3, m_4; g) = \alpha(m_1, m_2, m_3, m_4)d(m_1, m_2; g).
\]

Accordingly, we compared the divergence in proteins encoded by single-copy genes between *P. vivax* and *P. knowlesi* to protein divergence in four other *Plasmodium* species pairs (fig. 2; supplementary table S1, Supplementary Material online). Our pairwise approach overcomes problems posed by large differences in nucleotide composition between *Plasmodium* genomes and by incomplete genomes in some species. In particular, *P. vivax* and *P. knowlesi* have relatively high genomic GC content (>37% GC) compared with all other species in this study (<23% GC). Moreover, while species with completed genomes (e.g., *P. vivax*, *P. knowlesi*, or *P. falciparum*) have the full complement of greater than 5,000 protein-coding genes available, species with draft genomes (e.g., *P. reichenowi*) have only a small fraction (Materials and Methods). A pairwise approach allows individual data sets to include genes not shared by all species in our study. The divergence between species pairs was compared using total least squares (TLS) regressions. TLS has an important advantage over Least Squares in that it takes into account the error associated with all regression variables, so it downweights points with large uncertainties in their x or y coordinates (Materials and Methods).

The model described above provides an excellent fit to the data. An \( R^2 \)-statistic gave the fraction of the variation between lineages explained by pairwise correlations. The four data sets yielded values of \( R^2 \) from 47% to 83% (fig 3; Materials and Methods). The difference in magnitude of \( R^2 \) between the four analyses depended largely on the divergence between the species pairs shown on the x axis (note that the species pair in the y axis is the same for all analyses). In particular, larger median values of \( d_{AA} \) of the species pair on the x axis corresponded to larger \( R^2 \) values for the correlation (table 1 and fig. 3). At least three phenomena (and possibly all three together) explain these results. First, the stochastic variation in \( d_{AA} \) is relatively large for closely related species (being by rule of thumb roughly proportional to the square root of \( d_{AA} \)). Second, speciation events are often associated with population bottlenecks (Nei 1987; Hughes 2008), which lead to relatively high frequency of slightly deleterious substitutions. Finally, our methods detect most genes at small evolutionary distances, even those that are becoming pseudogenes, or evolving under diversifying selection, in one of the species. As genes accumulate indels, missense, and nonsense mutations our “HQ” data sets (Materials and Methods) fail to capture them.

The large \( R^2 \) values conclusively show that the relative rate of evolution of single-copy genes has remained remarkably constant across independent *Plasmodium* lineages, a necessary condition for the existence of protein-specific molecular clocks. Our pairwise approach does not test whether the overall rate of evolution is unvarying between lineages, the other requirement of a protein-specific clock. However, similar values of \( d_{ns}/d_{s} \), the ratio of nonsynonymous to synonymous substitution rates, in different *Plasmodium* lineages strongly suggests a constant evolution rate across the genus (Silva et al. 2011). Remarkably, the protein-specific clock applies over a wide range in evolutionary rates (fig. 3), a result crucial for phylogenetic studies across a range of divergence times (Bromham and Penny 2003). For each protein, the regression residual measures the protein’s conformity to a clock (supplementary table S1, Supplementary Material online).

Many *Plasmodium* genes evolving under strong diversifying selection will be absent from our data sets, because they will not satisfy the stringent homology criteria of our HQ data sets. In fact, of a set of 43 *P. falciparum* genes expected to evolve under positive selection (Weedall et al. 2008) only

---

**Table 1.** Protein Sequence Divergence Estimates.

| Pairwise Comparisons | \( G^2 \) | Median \( d_{AA} \) (Minimum–Maximum) |
|----------------------|----------|--------------------------------------|
| *P. vivax*–*P. knowlesi* | 2,820 | 0.179 (0.00001–2.62) |
| *P. falciparum*–*P. reichenowi* | 445 | 0.017 (0.00001–0.23) |
| *P. yoelii*–*P. berghei* | 761 | 0.053 (0.00001–4.04) |
| *P. yoelii*–*P. chabaudi* | 420 | 0.100 (0.00001–3.90) |
| *P. falciparum*–*P. yoelii* | 1,560 | 0.461 (0.00001–3.46) |

\(^{\text{a}}\text{Number of single-copy protein-coding genes in each pairwise comparison that satisfy the conditions that define HQ data sets (Materials and Methods).}\)
seven are captured in at least one of our four comparisons (supplementary table S2, Supplementary Material online). As expected for genes evolving under diversifying selection, $d_{AA}$ for these genes is relatively high, with all values above the 50th percentile. The regression residual values associated with these genes are also relatively high (none is among the lowest 25th percentile), suggesting that patterns of positive selection are not necessarily constant across lineages, as previously observed (Weedall et al. 2008).

Relative and Absolute Divergence Times

The slope of TLS regressions for the $P. vivax$ and $P. knowlesi$ pair with the four other species pairs specifies the relative ages of the corresponding four splits (fig. 3 and table 2). The age of the split between $P. vivax$ and $P. knowlesi$ is approximately 30% (or 0.285) of the age of the split between $P. falciparum$ and $P. yoelii$, and it was approximately 1.4, 2.6, and 6.1 times older than the split between $P. yoelii$–$P. berghei$, $P. berghei$–$P. chabaudi$, and $P. falciparum$–$P. reichenowi$, respectively. The most recent common ancestor of these seven mammalian parasites is represented by the node that gave rise to the lineages leading to $P. falciparum$ and to $P. yoelii$ (Silva et al. 2011; fig. 2). Accordingly, our results indicate that the sampled mammalian $Plasmodium$ parasites had a common origin about 22 times earlier than the split between the youngest species pair in this data set, $P. falciparum$ and $P. reichenowi$.

The absolute age of the split between $P. vivax$ and $P. knowlesi$ can be estimated by calibrating nucleotide divergence in synonymous sites between the two species (median $d_S = 0.55 \pm 0.0063$; [Carlton et al. 2008]) with the rate of synonymous substitution per site per year in metazoans, approximately $8.1 \times 10^{-9}$ (Lynch and Conery 2000), and in invertebrates, approximately $1.5 \times 10^{-8}$ (Li 1997; table 2). Accordingly, the split between $P. vivax$ and $P. knowlesi$ occurred an estimated 18–34 Ma, depending on the calibration rate used. The age of the split between $P. vivax$ and $P. knowlesi$ then yielded absolute ages of the split between remaining species pairs (table 2). According to these estimates, the age of the split between $P. falciparum$ and its sister species $P. reichenowi$ was about 3.0–5.5 Ma, rodent parasites diversified 13–25 Ma, and the sampled mammalian $Plasmodium$ parasites had a common origin at least 64 Ma.

We applied this novel approach to dating divergence times to two other taxonomic groups with extensive genomics resources, in particular placental mammals (Douzery et al. 2008).
The estimated divergence time between human and chimpanzee (Hobolth et al. 2007), and is congruent with the cospeciation of these *Plasmodium* species with their hosts (Escalante et al. 1995; Hughes and Verra 2010). However, a parasite split postdating the human–chimpanzee split, with the parasite switching hosts before the origin of anatomically modern humans, possibly in the late Pliocene (Martin et al. 2005), is also compatible with the results. If indeed *P. falciparum* is primarily a parasite of gorillas (Liu et al. 2010), its split from *P. reichenowi* would have postdated that of the hosts, which occurred greater than 8 Ma (Raam et al. 2005; Steiper and Young 2006; Yang and Rannala 2006). The corresponding sequences indicate, however, that the split between *P. falciparum* and *P. reichenowi* could not have occurred within the last 10,000 years (Rich et al. 2009), because a split so recent entails a rate of synonymous substitution in *Plasmodium* two orders of magnitude higher than any observed in eukaryotes. We also investigated the possibility that sequencing errors in the draft genome assembly of *P. reichenowi* inflated amino acid divergence relative to *P. falciparum*. We obtained all *P. reichenowi* sequences available in GenBank, and compared them to those in our preliminary annotation of the draft genome assembly (Materials and Methods). Of the 97 *P. reichenowi* protein sequences available, 41 had significant matches to 19 of the 698 *P. reichenowi* protein data set in our preliminary annotation (supplementary table S3, Supplementary Material online). These 19 unique protein sequences are nearly identical to the sequences inferred from the genome assembly (sequence identity: median = 100%; average = 98.9%). Of the four out of 19 proteins for which sequence identity was less than 99% when compared with the draft genome, one is a hypothetical protein, and two others are homologous to *P. falciparum* antigens, and are therefore expected to have a high degree of amino acid sequence polymorphism. Therefore, although the draft genome assembly of *P. reichenowi* probably contains sequencing errors, the protein sample available suggests that the errors could not be sufficiently frequent to alter the estimates of divergence between *P. falciparum* and *P. reichenowi* noticeably.

Previous estimates of the age of the split between *P. vivax* and *P. knowlesi*, which placed the speciation event within the past 7 My (Escalante et al. 1998, 2005; Jongwutiwes et al. 2005; Hayakawa et al. 2008), all used mitochondrial sequences.

### Table 2. Divergence Times of Mammalian *Plasmodium* Parasites.

| Four-Way Comparisons | G* | Divergence of Pv–Pk Relative to Other Species Pairsa (95% CI) | Age of Divergence of Each Species Pair, Calibrated with Two Eukaryotic Substitution Rates (My) |
|----------------------|----|----------------------------------------------------------|-----------------------------------------------------------------------------------|
|                      |    | Invertebrate Rate | Metazoan Rate |
| *P. vivax*–*P. knowlesi* | 280 | 6.13 (5.84–6.42) | 18.3b | 34.0b |
| *P. falciparum*–*P. reichenowi* | 497 | 2.64 (2.54–2.73) | 3.0 (2.9–3.1) | 5.5 (5.3–5.8) |
| *P. yoelii*–*P. berghei* | 260 | 1.37 (1.30–1.44) | 6.9 (6.7–7.2) | 12.9 (12.5–13.4) |
| *P. yoelii*–*P. chabaudi* | 1,018 | 0.285 (0.280–0.290) | 13.4 (12.7–14.1) | 24.8 (23.6–26.2) |
| *P. falciparum*–*P. yoelii* | 260 | 0.285 (0.280–0.290) | 64.2 (63.1–65.4) | 119.3 (117.2–121.4) |

*Number of single-copy protein-coding genes available for all four species in a comparison.

*Relative age of the most recent common ancestor (MRCA) of *P. vivax* and *P. knowlesi* (Pv–Pk), in relation to the age of the MRCA of other species pairs, obtained from the slope, α, of the TLS regression; the 95% confidence intervals (CI) were obtained by bootstrap (Materials and Methods).

*The absolute age of the MRCA of *P. vivax* and *P. knowlesi* was obtained by calibrating the median value of d, 0.0063; first quartile d = 0.39; third quartile d = 0.81) between the two species, obtained from 3,234 single-copy protein-coding genes (Carlton et al. 2008), with the average rate of evolution of synonymous sites in Drosophila (1.5 × 10–9/site/year; Li 1997) and in metazoa (8.1 × 10–9/site/year; Lynch and Conery 2000).
However, the mitochondrial genome in *Plasmodium* has properties that lead to systematic underestimation of sequence divergence (see below), and is therefore unsuitable for dating. Our estimates, based on thousands of nuclear genes, place this split much earlier in time, between 18 and 34 Ma. Interestingly, the estimates overlap with the split of their respective hosts, apes (Hominioidea), and Old World monkeys (Cercopithecidae), about 24–38 Ma (Steiper and Young 2009; fig. 4). Most significantly, recent studies strongly suggest that in fact *P. vivax* originated in Africa, and that its closest taxa are parasites of chimpanzees and gorillas (Liu et al. 2014).

The rodent parasites included here, *P. yoelii*, *P. chabaudi*, and *P. berghei*, all parasitize murine rodents. The most recent common ancestor of the parasites lived between 13 and 25 Ma, suggesting that the diversification of these rodent parasites coincided with the radiation of the family Muridae, which occurred in the last 25 Ma (Steppan et al. 2004).

Finally, our analyses place the origin of the mammalian *Plasmodium* in the late Mesozoic, between 64 and 120 Ma, an interval overlapping with the divergence between the primate and rodent lineages (Steppan et al. 2004; Murphy and Eizirik 2009), strongly suggesting that mammals and *Plasmodium* have coexisted for much, and possibly all, of their evolutionary history (fig. 4). Our results agree with studies placing the ages of the *P. falciparum–P. reichenowi* and *P. vivax–P. knowlesi* speciation events toward the older end of previous range estimates.

**Discussion**

The rate of substitution in *Plasmodium* is unknown, precluding reliable age estimates for the most recent common ancestor of extant human parasite populations and for speciation events within the genus (Prugnolle et al. 2011). Instead, published ages are often obtained by calibrating DNA sequence polymorphism (usually in *P. falciparum* or in *P. vivax* [Joy et al. 2003; Escalante et al. 2005; Jongwutiwes et al. 2005; Mu et al. 2005]) or sequence divergence between species (Hayakawa et al. 2008; Ricklefs and Outlaw 2010; Pacheco et al. 2012) with a substitution rate inferred under the assumption of cospeciation of parasite species and their respective host. Examples include the cospeciation of *P. falciparum* and *P. reichenowi* with human and chimpanzees (Joy et al. 2003; Jongwutiwes et al. 2005), the radiation of monkey parasites being coincident with that of their respective Old World monkey host species (Escalante et al. 2005; Mu et al. 2005; Hayakawa et al. 2008; Krief et al. 2010; Pacheco et al. 2012), or the cospeciation of avian species pairs and their respective *Plasmodium* parasites (Ricklefs and Outlaw 2010). Other studies have calibrated divergence with rRNA substitution rates estimated for bacterial or eukaryotic taxa (Escalante and Ayala 1995; Escalante et al. 1995). All these studies share the common problem of examining at most a few genetic loci, whose polymorphism or divergence might be skewed in an unknown but specific manner.

In 2011, Silva et al. (2011) obtained times for several speciation events among mammalian *Plasmodium* parasites by applying several methods to sequences from 45 nuclear loci. However, their study assumed that *P. falciparum* and *P. reichenowi* cospeciated with their respective mammalian hosts, which may not be accurate (Liu et al. 2010).

Here, we approached the problem differently. *P. vivax* and *P. knowlesi* have virtually closed genome assemblies and are sufficiently closely related to avoid saturation of synonymous sites. Over 2,800 protein-coding nuclear genes can be reliably aligned between these species in an automated fashion, resulting in a highly accurate estimate of synonymous substitutions rate per site between the two species ($d_S$ (Py-Pk)). We calibrated $d_S$ (Py-Pk) with two eukaryotic rates of synonymous substitution per site per year, one obtained from *Drosophila* (Li 1997) and one a rough average value for metazoans (Lynch and Conery 2000), to obtain an estimate of the divergence time between the two *Plasmodium* species. Because the evolution rate per year in *Drosophila*
in the et al. 2010), bringing to seven the number of identified species Plasmodium the most deadly of the human Duval et al. 2010; Krief et al. 2010; Liu et al. 2010; Prugnolle unusually high rates of molecular evolution, host switching, and health perspective, the mitochondrial estimates imply provide context for the biology and evolution of crucial finding for the study of malaria, as these species will observed for yeast, Drosophila and humans (Bopp et al. 2013), although it remains unclear how this translates to substitution rates. It is also noteworthy that the closest eukaryotic taxon to the Apicomplexa with a significant fossil record, the diatoms, have an estimated $d_0 \sim 6.5 \times 10^{-9}$ (Sorhannus and Fox 1999), which overlaps with the more conservative metazoan rate estimate. Finally, we had the relative divergence time between species pairs from the correlations of hundreds to thousands of genes, so the divergence times between Plasmodium and Plasmodium permitted us to estimate the corresponding absolute divergence times.

Five new putative species of Plasmodium closely related to P. falciparum and P. reichenowi have recently been discovered, all of which infect chimpanzees or gorillas (Ollomo et al. 2009; Duval et al. 2010; Krief et al. 2010; Liu et al. 2010; Prugnolle et al. 2010), bringing to seven the number of identified species in the Laverania subgenus (Prugnolle et al. 2011). This is a crucial finding for the study of malaria, as these species will provide context for the biology and evolution of P. falciparum, the most deadly of the human Plasmodium parasites. So far only data for mitochondrial loci have been published, but soon the defining characteristics of each species will be identified from the nuclear genome, such as differences in genome sequence and structure, gene composition, and sequence divergence. Of particular interest will be the evolution of genes involved in adhesion and invasion of the host cell, and those responsible for evasion of the host immune system. The age of the species will provide the defining rate of evolution of these pathogenesis traits. The mitochondrial data suggest that the common origin of the seven great ape parasites is roughly two to three times older than the split of P. falciparum from P. reichenowi (Prugnolle et al. 2011), although data from many nuclear loci will be necessary to determine this with higher accuracy. Based on our estimate of the P. falciparum–P. reichenowi split, the mitochondrial data place the origin of the Laverania subgenus at approximately 9–16 My old. This suggests that the diversification of these parasites is contemporaneous with that of the Homininae subfamily (containing humans, chimpanzees, bonobos, and gorillas), which took place in the past 15 My (Raoult et al. 2005; Yang and Rannala 2006; Horner et al. 2007).

A recent study based on the mitochondrial gene cytochrome b estimated the split between P. falciparum and P. reichenowi at 2.5 Ma and the origin of all mammalian Plasmodium at less than 13 Ma (Ricklefs and Outlaw 2010). If the mitochondrial estimates are correct, the greater part of mammalian evolution must have occurred in a context devoid of Plasmodium. More importantly from a human health perspective, the mitochondrial estimates imply unusually high rates of molecular evolution, host switching, and speciation as defining characteristics of Plasmodium. However, estimates based on mitochondrial data may be unreliable because the Plasmodium mitochondrial genome has an extremely biased nucleotide composition (27.2% GC overall, 13.7% GC in third codon positions) and a high saturation rate (McIntosh et al. 1998). Our results support this conclusion by showing that for very recent divergences, where substitution saturation is rare, our study generally confirmed the age estimates (e.g., age of P. falciparum–P. reichenowi split), but that for older divergences there were sharp differences, such as the greater than 5-fold difference for the origin of mammalian Plasmodium. On the other hand, if the mitochondrial estimates are correct, they imply that the amino acid substitution rate in Plasmodium nuclear proteins decreased sharply through time, a phenomenon for which there is no seeming explanation.

Overall, our results indicate that mammalian-infecting Plasmodium species have broadly coexisted, and perhaps coevolved, with their mammalian hosts. Even though the transmission of Plasmodium between humans and other primates has been documented (Cox-Singh et al. 2008; Krief et al. 2010; Liu et al. 2010), there is no evidence that this was a common event on an evolutionary scale (Wanaguru et al. 2013).

There is evidence that malaria has had a substantial impact on recent human evolution (Kwiatkowski 2005), and conversely that Plasmodium is under strong pressure exerted by its host (Mackinnon and Marsh 2010). Our results strengthen the hypothesis that the mammalian lineage carried malaria parasites long before the infection of humans, as suggested by the rapid coevolution of surface glycoproteins on red blood cells and their Plasmodium-encoded ligands (Wang et al. 2003). Interestingly, the two best-studied mammalian immune systems, those of human and the house mouse (Mus musculus), have many known differences, several of which involve defense mechanisms implicated in the response to malaria (Coban et al. 2007), including the differential expansion and deletion of various sets of natural killer cell receptors (Hao et al. 2006) and mannose-binding lectins (Sastry et al. 1995), differences between the roles of CD3y and CD3g (Fernandez-Malave et al. 2006), and differences in the structure of the splenic marginal zone (Steiniger et al. 2005). Thus, some of the distinctive immune mechanisms in different mammalian lineages may reflect a lengthy coevolution with a unique set of malaria parasites.

Materials and Methods

Data

The genomic files for the seven Plasmodium species were downloaded from PlasmoDB 5.5. GFF files describing gene, protein, coding sequences (CDS) and exon features were also downloaded for all six annotated species, P. berghei (Pb), P. chabaudi (Pc), P. falciparum (Pf), P. knowlesi (Pk), P. vivax (Pv), and P. yoelii (Py). Protein sequences and CDS chromosome coordinates were compared using in-house scripts, and reconciled when possible. The P. reichenowi (Pr) genome did not have an associated annotation, and a putative gene set was derived based on the P. falciparum gene set.
We used PASA (Haas et al. 2003) to generate a gene set for *P. reichenowi*, using the *P. falciparum* gene set as a proxy for full-length transcript data, and requiring ≥90% nucleotide identity over ≥90% of the length of the gene. This recovered 698 putative proteins. We defined COGs (clusters of orthologous genes) based on our comparative pipeline, which starts with BLASTP to find matches among protein sequences within and across species. We used the BLOSUM62 matrix with expected value 10\(^{-5}\). Jaccard clustering was then performed twice, once to form within-species clusters of paralogous genes and a second time to derive a set of multispecies COGs. In the first case, we used an 80% identity cutoff and a link score of 0.6. In the second, we set a Jaccard coefficient cutoff of 0 for edge pruning. Genes within a COG were aligned with ClustalW with default parameters.

For each of the five pairwise data sets, or lineages (fig. 2), we retained only those genes satisfying the following conditions: 1) the genes have no paralogs in either species; 2) the genes and respective proteins have no irreconcilable differences stemming from incorrect annotation (see previous paragraph); and 3) the protein sequence lengths in the two species are within 10% of each other. The conditions minimize errors in estimated evolutionary rates due to paralogy, incorrect inferences of orthology, annotation errors, or gap-induced misalignments. The resulting five HQ data sets contained the following numbers of genes successfully paired between the corresponding species: *Pv*- *Pb*, 2827; *Py*– *Pc*, 853; *Py*– *Pb*, 1,220; *Pf*- *Pr*, 454; and *Pf*- *Py*, 1,400. Table 2 displays results for four regressions using genes present in both the *Pv*- *Pb* HQ data set and each of the other HQ data sets, in turn.

Molecular Evolution

Amino acid sequence divergence (d\(_{AA}\)) was estimated using the JTT-F model of evolution implemented in the program codeml from the PAML package (Yang 2007). IDEA (Egan et al. 2008) was used to launch and monitor the PAML analyses and to distribute the computations across a grid of computers.

TLS Regression

Unlike the true amino acid sequence distances \(d(m_1, m_2; g)\) and \(d(m_3, m_4; g)\), the observed distances \(D(m_1, m_2; g)\) and \(D(m_3, m_4; g)\) contain noise, because the corresponding mutational processes are random. The taxon pairs \((m_1, m_2)\) and \((m_3, m_4)\) were chosen to correspond to separate tree-paths in the phylogenetic tree, thereby justifying the assumption that the two random mutational processes differentiating each taxon pair are independent. To simplify the notation and to prepare for a regression, fix the malarial species \(m_1, m_2, m_3, \) and \(m_4\), so \(x = \alpha (m_1, m_2, m_3, m_4)\) is a constant. Let \(X_g = D(m_1, m_2; g)\) and \(Y_g = D(m_3, m_4; g)\) denote the estimated sequence distances, ranging over the genes \(g\) common to \(m_1, m_2, m_3, \) and \(m_4\). Similarly, let \(x_g = d(m_1, m_2; g)\) and \(y_g = d(m_3, m_4; g)\) denote the true sequence distances. Our null hypothesis rests on the assumption that \(y_g = \alpha x_g\) in equation (3). Under its evolutionary models, PAML (Yang 2007) estimates the amino acid sequence divergences \((X_g, Y_g)\) with maximum likelihood (ML), deriving the corresponding error estimates \((s_{x_g}, s_{y_g})\) from observed Fisher information (Kendall and Stuart 1979). We applied TLS regression to find \(\{x_g\}\) and \(\alpha\) minimizing the sum

\[
\sum_{g=1}^{G} \left[ \frac{(X_g - x_g)^2}{2s_{x_g}^2} + \frac{(Y_g - \alpha x_g)^2}{2s_{y_g}^2} \right],
\]

where the sum is restricted to those genes \(g = 1, \ldots, G\) satisfying the cutoffs \(X_g/s_{x_g} \geq 2\) and \(Y_g/s_{y_g} \geq 2\). A confidence interval for \(\alpha\) was then obtained by bootstrapping the residuals \((\hat{x}_g - X_g)/s_{x_g}\) and \((\hat{y}_g - Y_g)/s_{y_g}\).

Derivation of the TLS Regression in Equation (4)

PAML produces its ML estimation (MLE) \(X_g = D(m_1, m_2; g)\) of a sequence distance \(x_g = d(m_1, m_2; g)\) from a log-likelihood \(\lambda(X_g, x_g)\), which corresponds to a particular random model of evolution. Under our null model \(y_g = \alpha x_g\), the evolutionary model also specifies \(\lambda(Y_g | y_g) = \lambda(Y_g | \alpha x_g)\). Thus, the log-likelihood of the pertinent data is

\[
\lambda(X, Y; \alpha) = \sum_{g} [\lambda(X_g | x_g) + \lambda(Y_g | \alpha x_g)],
\]

as a function of \(\alpha\), where the sum is over all genes \(g\) common to \(m_1, m_2, m_3, \) and \(m_4\). The likelihood in equation (5) specifies our statistical null model completely.

Unfortunately, PAML does not evaluate the log-likelihoods \(\lambda(X_g | x_g)\) and \(\lambda(Y_g | y_g) = \lambda(Y_g | \alpha x_g)\). Instead, PAML reports MLEs \((X_g, Y_g)\) maximizing the individual terms \(\lambda(X_g | x_g)\) and \(\lambda(Y_g | y_g)\), along with error estimates \((s_{x_g}, s_{y_g})\) derived from observed Fisher information (Kendall and Stuart 1979).

To exploit this reduced information as much as possible, approximate \(X_g\) as a normal variate with its mean \(x_g\) unknown and known standard deviation \(s_{x_g}\), so

\[
\tilde{\lambda}(X_g | x_g) \approx \lambda(X_g | x_g) = \frac{1}{2} \ln(2\pi) - (X_g - x_g)^2 / (2s_{x_g}^2),
\]

and similarly for \(Y_g\). The substitution of the quadratic approximation \(\tilde{\lambda}\) for \(\lambda\) in equation (5) leads to the problem of maximizing

\[
\tilde{\lambda}(X, Y; \alpha) = \sum_{g} \left[ \frac{(X_g - x_g)^2}{2s_{x_g}^2} + \frac{(Y_g - \alpha x_g)^2}{2s_{y_g}^2} \right],
\]

where the sum is over all genes \(g\) common to \(m_1, m_2, m_3, \) and \(m_4\), a maximization almost equivalent to the TLS regression in equation (4). The rates of evolution \(x_g\) and \(y_g = \alpha x_g\) must be positive; however, so in practice, the normal approximation \(\tilde{\lambda}(X, Y; \alpha)\) becomes problematic when it puts appreciable probability mass on negative values, skewing the TLS fit.

For any gene \(g\), and for \(X_g < 0\) or \(Y_g < 0\), the true (unknown) likelihoods \(\lambda(X_g | x_g)\) and \(\lambda(Y_g | y_g)\) are 0. The normal approximation therefore overestimates the true...
likelihood near \( X_{g} = 0 \) or \( Y_{g} = 0 \). Hence, if \( X_{g}/s_{X_{g}} \) or \( Y_{g}/s_{Y_{g}} \) is small, the normal approximation might be inaccurate. In addition, from a scientific perspective, the gene \( g \) is of little interest, because it might not have evolved much. It therefore seems advisable to exclude \( g \) from the TLS fit. The \( R^{2} \)-statistic, the estimated fraction of the variation in the data that the regression explains, is then a conservative underestimate.

Accordingly, we applied a cutoff \( X_{g}/s_{X_{g}} \geq 2 \) or \( Y_{g}/s_{Y_{g}} \geq 2 \), dropping any gene \( g \) where the normal approximation in equation (7) has more than 5% of either marginal probability on the negative numbers, arriving at the TLS regression in equation (4).

The bootstrap yields an approximate confidence interval for \( \alpha \), as follows. Under the null hypothesis, the normalized residuals \( r_{x_{g}} = (X_{g} - \hat{X}_{g})/s_{X_{g}} \) and \( r_{y_{g}} = (Y_{g} - \hat{Y}_{g})/s_{Y_{g}} \) can be bootstrapped to approximate the variance of \( \alpha \) in the original sample, as follows. Let \( \pi \) and \( \hat{\pi} \) denote permutations of the \( G \) usable genes (the genes \( g = 1, \ldots, G \) with \( X_{g}/s_{X_{g}} \geq 2 \) and \( Y_{g}/s_{Y_{g}} \geq 2 \)). For 10,000 uniformly random pairs of independent permutations \( \pi \) and \( \hat{\pi} \), we calculated a TLS regression slope \( \alpha^{*} \) by bootstrapping the residuals \( r_{x_{g}} \) and \( r_{y_{g}} \) to produce values \( X_{g}^{\hat{\pi}} = x_{g} + r_{x_{g}} \hat{\pi} \) and \( Y_{g}^{\pi} = y_{g} + r_{y_{g}} \pi \). The normal approximation and the bootstrapped sample standard deviation \( s_{g}^{\hat{\pi}} \) of the values \( \alpha^{*} \) gave an estimated 95% confidence interval \( [\alpha - z_{0.025} s_{g}^{\hat{\pi}}, \alpha + z_{0.025} s_{g}^{\hat{\pi}}] \) for \( \alpha \).

The Fraction of Variation in the Data Explained by TLS Regression

To assess the extent to which \( P. falciparum \) amino acid substitution rates conform to a molecular clock model, we quantify the strength of the relationship between substitution rates in the various lineages. This can be measured by quantifying the strength of the relationship between substitution rates conform to a molecular clock model, we compute the \( R^{2} \)-statistic \( \hat{\rho}_{\text{att}} \), which corresponds to the fraction of the variance in one variable explained by dependence on the other variable. This method was implemented in R. Source code available from the authors upon request.

Supplementary Material

Supplementary table S1–S3 and methods S1 are available at Molecular Biology and Evolution online (http://www.mbe.oxfordjournals.org/).

Acknowledgments

The authors thank Austin Hughes for stimulating discussions, and J. Crabtree, A. Ganapathy, and D. Riley for technical support. This project was funded in part with federal funds from the National Institute of Allergy and Infectious Diseases (NIAID), National Institutes of Health (NIH), Department of Health and Human Services under contract number HHSN272200900009C and by startup funds to J.C.S., and by the Intramural Research Program of the National Institutes of Health, National Library of Medicine (J.L.S.).

References

Bopp SE, Manary MJ, Bright AT, Johnston GL, Dharia NV, Luna FL, McCormack S, Plouffe D, McNamara CW, Walker JR, et al. 2013. Mitotic evolution of \( P. falciparum \) shows a stable core genome but recombination in antigenic families. PLoS Genet. 9: e1003293.

Bromham L, Penny D. 2003. The modern molecular clock. Nat Rev Genet. 4:216–224.

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 \( P. vivax \). Nature 455:757–763.

Carlton JM, Angiuoli SV, Suh BB, Kooij TW, Perea 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 \( P. yoelii \). Nature 419:512–519.

Coban C, Ishii KJ, Horii T, Akira S. 2007. Manipulation of host innate immune responses by the malaria parasite. Trends Microbiol. 15: 271–278.

Cox-Singh J, Davis TM, Lee KS, Shamsul SS, Matusop A, Ratnam S, Rahman HA, Conway DJ, Singh B. 2008. \( P. knowlesi \) malaria in humans is widely distributed and potentially life threatening. Clin Infect Dis. 46:165–171.

Douzerel EJ, Scornavacca C, Romiguier J, Belkhir K, Galtier N, Deluc F, Runzewicz V. 2014. OrthoMAM v8: a database of orthologous exons and coding sequences for comparative genomics in mammals. Mol Biol Evol. 31(7):1923–1928.

Duval L, Fourment M, Merrien E, Rouset D, Sadeuh SA, Goodman SM, Andriaolininina NV, Randrianarivelojosia M, Paul RE, Robert V, et al. 2010. African apes as reservoirs of \( P. falciparum \) and the origin and diversification of the Laverania subgenus. Proc Natl Acad Sci U S A. 107:10561–10566.

Egan A, Maharuk A, Crabtree J, Badger JH, Carlton JM, Silva JC. 2008. IDEA: interactive display for evolutionary analyses. BMC Bioinformatics 9:524.
and Mbl2 reveals features common to other collectin genes. *Mamm Genome.* 6:103–110.

Silva JC, Egan A, Friedman R, Munro JB, Carlton JM, Hughes AL. 2011. Genome sequences reveal divergence times of malaria parasite lineages. *Parasitology* 138(13):1737–1749.

Sokal RR, Rohlf FJ. 1981. Biometry: the principles and practice of statistics in biological research. San Francisco: W. H. Freeman.

Sorhannus U, Fox M. 1999. Synonymous and nonsynonymous substitution rates in diatoms: a comparison between chloroplast and nuclear genes. *J Mol Evol.* 48:209–212.

Steiniger B, Timphus EM, Jacob R, Barth PJ. 2005. CD27+ B cells in human lymphatic organs: re-evaluating the splenic marginal zone. *Immunology* 116:429–442.

Steiper ME, Young NM. 2006. Primate molecular divergence dates. *Mol Phylogenet Evol.* 41:384–394.

Steiper ME, Young NM. 2009. Primates (Primates). In: Hedges SB, Kumar S, editors. The Timetree of Life. New York: Oxford University Press. p. 482–486.

Steppan S, Adkins R, Anderson J. 2004. Phylogeny and divergence-date estimates of rapid radiations in muroid rodents based on multiple nuclear genes. *Syst Biol.* 53:333–553.

Tanabe K, Mitaka T, Jombart T, Eriksson A, Horibe S, Palacpac N, Ranford-Cartwright L, Sawai H, Sakihama N, Ohmoe H, et al. 2010. *Plasmodium falciparum* accompanied the human expansion out of Africa. *Curr Biol.* 20:1–7.

Throckmorton LH. 1975. The phylogeny, ecology, and geography of *Drosophila*. In: King RC, editor. Handbook of genetics. New York: Plenum. p. 421–469.

Valkiunas G, Ashford RW, Bensch S, Killick-Kendrick R, Perkins S. 2011. A cautionary note concerning *Plasmodium* in apes. *Trends Parasitol.* 27:231–232.

Wanaguru M, Liu W, Hahn BH, Rayner JC, Wright GJ. 2013. RH5-Basigin interaction plays a major role in the host tropism of *Plasmodium falciparum*. *Proc Natl Acad Sci U S A.* 110:20735–20740.

Wang HY, Tang H, Shen CK, Wu CI. 2003. Rapidly evolving genes in human. I. The glycoporphins and their possible role in evading malaria parasites. *Mol Biol Evol.* 20:1795–1804.

Weedall GD, Polley SD, Conway DJ. 2008. Gene-specific signatures of elevated non-synonymous substitution rates correlate poorly across the *Plasmodium* genus. *PLoS One* 3:e2281.

World Health Organization. 2009. World malaria report 2009. Geneva: World Health Organization.

Yang Z. 2007. PAML 4: phylogenetic analysis by maximum likelihood. *Mol Biol Evol.* 24:1586–1591.

Yang Z, Rannala B. 2006. Bayesian estimation of species divergence times under a molecular clock using multiple fossil calibrations with soft bounds. *Mol Biol Evol.* 23:212–226.

Zuckerkandl E, Pauling L. 1965. Evolutionary divergence and convergence in proteins. In: Bryson V, Vogel HJ, editors. Evolving genes and proteins. New York: Academic Press. p. 97–166.