

Evolution of Noncoding and Silent Coding Sites in the *Plasmodium falciparum* and *Plasmodium reichenowi* Genomes

Daniel E. Neafsey,*¹ Daniel L. Hartl,* and Matt Berriman†

*Department of Organismic and Evolutionary Biology, Harvard University, Cambridge, MA; and

†Wellcome Trust Sanger Institute, Hinxton, United Kingdom

We compared levels of sequence divergence between fourfold synonymous coding sites and noncoding sites from the intergenic and intronic regions of the *Plasmodium falciparum* and *Plasmodium reichenowi* genomes. We observed significant differences in the level of divergence between these classes of silent sites. Fourfold synonymous coding sites exhibited the highest level of sequence divergence, followed by introns, and then intergenic sequences. This pattern of relative divergence rates has been observed in primate genomes but was unexpected in *Plasmodium* due to a paucity of variation at silent sites in *P. falciparum* and the corollary hypothesis that silent sites in this genome may be subject to atypical selective constraints. Exclusion of hypermutable CpG dinucleotides reduces the divergence level of synonymous coding sites to that of intergenic sites but does not diminish the significantly higher divergence level of introns relative to intergenic sites. A greater than expected incidence of CpG dinucleotides in intergenic regions less than 500 bp from genes may indicate selective maintenance of regulatory motifs containing CpGs. Divergence rates of different classes of silent sites in these *Plasmodium* genomes are determined by a combination of mutational and selective pressures.

Introduction

The heterogeneous distribution of nucleotide diversity in the genome of the human malaria parasite *Plasmodium falciparum* is one of the most curious features to emerge from the genetic study of this organism. Many genes, particularly membrane proteins involved with drug resistance or antigenic variation (Verra and Hughes 2000; Polley and Conway 2001; Volkman et al. 2002), show high levels of synonymous and nonsynonymous polymorphism. For example, Hughes (1999) estimates that different allelic classes of the antigenic *msp1* gene are so divergent that they may have diverged over 48 MYA. Other genes in the *P. falciparum* genome exhibit nonsynonymous polymorphism but a complete lack of synonymous variation (Rich et al. 1998). Population genetic surveys of *P. falciparum* introns (Volkman et al. 2001) and mitochondrial genomes (Conway et al. 2000) indicate a dramatic paucity of noncoding and synonymous coding variations relative to the high level of nonsynonymous variation observed in a select set of antigenic genes.

There has been much debate regarding the cause of this nonuniform distribution of population genetic variation in the *P. falciparum* genome (see Hughes and Verra 2002; Hartl 2004, for review). The “Malaria’s Eve” hypothesis suggests that a population bottleneck which occurred approximately 10,000 years ago wiped away much of the polymorphism from the *P. falciparum* genome, with balancing selection and strong directional selection from the human immune system sustaining variation in a few select genes (Rich and Ayala 1998, 2000; Rich et al. 1998, 2000; Ayala, Escalante, and Rich 1999). Hypotheses favoring an ancient common ancestor (Hughes and Verra 1998, 2001) invoke selective constraint or some other variation-limiting

mechanism acting at silent coding and noncoding positions (Saul and Battistutta, 1988; Arnot 1991; Pizzi and Frontali 1999, 2000; Saul 1999; Forsdyke 2002; Jongwutiwes et al. 2002).

Forsdyke (2002) suggests that noncoding and silent sites in the *P. falciparum* genome may be especially subject to nonclassical selective pressures due to the high AT content (>80%) of its genome. He reports a correlation between local base order-dependent fold stability and intron positions in several *P. falciparum* genes, suggesting that noncoding sequences may be under selection to preserve the capacity of DNA or mRNA transcripts to form secondary structures. These observations corroborate the work of Pizzi and Frontali (1999, 2000), who detected evidence of conservation of nucleotide composition and certain oligonucleotide motifs when aligning *P. falciparum* introns to orthologs from the rodent parasite *Plasmodium berghei*.

It is therefore of considerable interest to understand the mutational and selective pressures acting on noncoding and silent sites in *P. falciparum*. To investigate the evolutionary history of these sequence classes, we used the chimpanzee parasite *Plasmodium reichenowi*, which is thought to have diverged from *P. falciparum* 6–10 MYA (Escalante and Ayala 1994) and is at an appropriate evolutionary distance for alignment of noncoding regions with *P. falciparum*. The *P. reichenowi* genome has been partially sequenced to approximately onefold coverage and assembled. The assembled contiguous sequences (contigs) cover slightly less than one-third of the *P. reichenowi* genome but provide ample intergenic, intronic, and silent coding sites for the analysis of divergence.

In comparing the level of sequence divergence between *P. falciparum* and *P. reichenowi*, we report that fourfold synonymous coding sites exhibit a level of divergence significantly higher than that of introns, which are in turn significantly more divergent than intergenic sequences. This pattern of divergence is most likely not due to annotation errors, as a similar pattern is observed in a more conservative consensus annotation. Exclusion of hypermutable CpG dinucleotides from the analysis removes the distinction in divergence rate between fourfold synonymous coding sites

¹ Present address: Broad Institute of MIT and Harvard, Cambridge, MA.

Key words: malaria, plasmodium, *reichenowi*, noncoding, divergence, introns.

E-mail: neafsey@broad.mit.edu.

Mol. Biol. Evol. 22(7):1621–1626. 2005

doi:10.1093/molbev/msi154

Advance Access publication April 27, 2005

and intronic sites, but intronic sites are still found to be significantly more divergent than nontranscribed intergenic sequences. We detect a greater than expected incidence of CpG dinucleotides in intergenic regions that are proximal to genes, indicating possible selective maintenance of CpG-containing regulatory motifs in these regions. We conclude that differences in the divergence rate of different noncoding and silent coding sites in *P. falciparum* are due to a combination of selective and mutational pressures.

Methods

The genome of *P. reichenowi* is being sequenced using whole-genome shotgun at the Wellcome Trust Sanger Institute (http://www.sanger.ac.uk/Projects/P_reichenowi). A set of 2,478 *P. reichenowi* genomic contigs (release 2003.01.23) was downloaded from PlasmoDB release 4.2 (<http://www.plasmodb.org>). The finished genomic scaffolds of *P. falciparum* (release 2002.10.03) were also downloaded from PlasmoDB. All *P. reichenowi* contigs were screened against the *P. falciparum* genomic scaffolds using Blast (Altschul et al. 1990), with a similarity cutoff criterion of $E < 1 \times 10^{-7}$. *Plasmodium reichenowi* contigs that failed to match any *P. falciparum* scaffolds or matched more than one scaffold were removed, resulting in a set of 1,537 *P. reichenowi* contigs for which orthologous *P. falciparum* genomic sequences could be confidently identified.

Matching regions identified during Blast analysis were aligned using ClustalW (Thompson, Higgins, and Gibson 1994) under default parameters. Intergenic, intronic, and coding sequences were identified using the annotation accompanying the official *P. falciparum* genome release (Gardner et al. 2002), which was obtained from PlasmoDB. Sequences were also divided into intergenic and intronic classes according to a more conservative consensus annotation created using only those nucleotide sites identified as intergenic or intronic by the release annotation, as well as predictions by Glimmer M (Delcher et al. 1999), Genefinder (Phil Green, <http://ftp.genome.washington.edu/cgi-bin/Genefinder>), and Phat (Cawley, Wirth, and Speed 2001).

Intergenic sequence fragments were divided into two categories to compare sequence divergence in regions that may be transcribed as 5' or 3' UTRs to intergenic sequences that are likely not part of genic transcripts. We designated all intergenic nucleotide positions within 500 bp of a coding region as "proximal intergenic" sites. Very few *P. falciparum* UTRs have been annotated, but Watanabe et al. (2002) conclude from the analysis of a complementary DNA library that the 5' UTRs of *P. falciparum* genes are unusually long, averaging 346 bp. Golightly et al. (2000) report a 3' UTR of 450 bp in an ookinete protein of the avian parasite *Plasmodium gallinaceum*. The proximal intergenic sequence class therefore likely captures most or all of the length of actual UTRs. Intergenic sites greater than 500 bp from a coding sequence were designated as "deep intergenic" sites.

Divergence estimates were calculated for each sequence class fragment according to the number of differences observed in a given sequence length. More complex nucleotide substitution models that correct for multiple hits were not employed due to their questionable ability to cope

with the highly skewed compositional bias of the genomes under consideration and the negligible effect they would have on the low divergence levels we observed. Gaps and ambiguous nucleotides were not included in divergence estimates. For coding regions, only fourfold synonymous third-position sites were analyzed due to the high level of codon bias in the 80% AT-rich *P. falciparum* and *P. reichenowi* genomes. The reading frame of coding regions was determined by realigning the *P. reichenowi* coding regions with a spliced coding sequence transcript of the matching *P. falciparum* gene recovered using Blast. Aligned coding regions that contained internal stop codons in either the *P. falciparum* or the *P. reichenowi* sequence were not considered. Overall divergence estimates for each class were obtained by dividing the total number of differences observed by the total length of aligned sequence fragments in the class. We determined 95% confidence intervals (CI) for the divergence estimate of each sequence class by a nonparametric bootstrapping process. For a given sequence class data set composed of n fragments, 1,000 new data sets of size n were constructed by random sampling of the fragments with replacement. Overall divergence estimates were then calculated for each bootstrap replicate and used to generate a distribution from which CI could be measured. The statistical significance of differences in divergence values for two sequence classes was determined by randomly pairing 1,000 bootstrap replicates from each class and observing the proportion of pairs in which one class is greater or less than the other.

We estimated the overall GC content of the different sequence classes, as well as the incidence of CpG dinucleotides, to determine whether these parameters correlate with divergence levels. The incidence of CpG dinucleotides was calculated for noncoding sequences as the number of observed CpG dinucleotides pairs divided by the total number of overlapping dinucleotide positions in the sequence ($[\text{sequence length in base pairs}] - 1$). The incidence of CpG dinucleotides in fourfold synonymous coding sites was calculated as the sum of the incidence of cytosine residues at fourfold sites that are followed by guanine residues in position 1 of the adjacent codon in either species, plus the incidence of guanine residues at fourfold sites that are preceded by a cytosine in position 2 of the same codon in either species, divided by the total number of aligned fourfold synonymous sites.

Divergence estimates were also calculated for all sequence classes after the removal of CpG dinucleotides. Nucleotide positions were excluded from divergence estimations if a CpG dinucleotide was detected in one or both of the genomes under comparison. The observed incidence of CpG dinucleotides in each sequence class was compared with an expectation of CpG incidence derived from the overall frequency of cytosine and guanine nucleotides in each class. For coding sequence, the expected incidence of CpG dinucleotides at fourfold synonymous third-position sites was calculated by adding the expected incidence of CpGs occupying codon positions 2 and 3 with the expected CpG incidence at codon positions 3 and 1, with guanine and cytosine frequencies independently estimated for each codon position. A chi-square test was used to calculate the significance of deviations from the expectation.

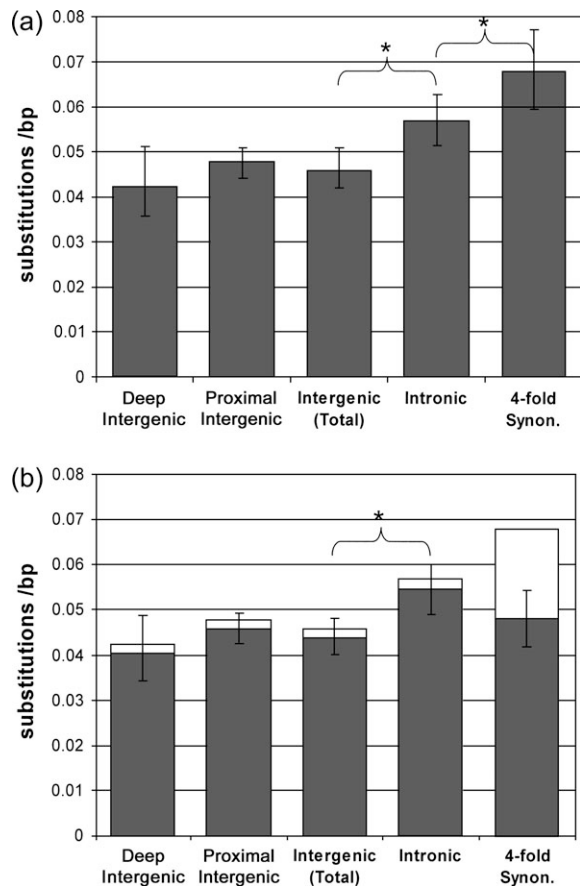


FIG. 1.—(a) Raw divergence estimates for different classes of noncoding and fourfold synonymous coding sites in the *Plasmodium falciparum* and *Plasmodium reichenowi* genomes. The “deep intergenic” and “proximal intergenic” sequence classes are subsets of the “intergenic (total)” sequence class. Error bars indicate 95% CI determined through a nonparametric bootstrapping process with 1,000 replicates. Bracketed pairs present significantly different divergence estimates, as determined by 1,000 replicates of a random bootstrap replicate pairing procedure described in *Methods*. (b) Raw divergence estimates for different classes of noncoding and fourfold synonymous coding sites in the *P. falciparum* and *P. reichenowi* genomes, with CpG dinucleotide sites excluded. Heights of shaded and unshaded bars, respectively, represent divergence calculations excluding and including CpG sites.

Results and Discussion

Overall *P. falciparum*/*P. reichenowi* divergence estimates for the different classes of noncoding sites and silent coding sites under the release annotation are presented in figure 1a. Under the release annotation, we recovered 1,076,514 bp of alignable intergenic sequence, 226,263 bp of alignable intronic sequence, and 2,265,114 bp of alignable coding sequence. Fourfold synonymous coding sites exhibit the highest level of divergence (0.068 substitutions/bp, 95% CI [0.060–0.077]), followed by introns (0.057 substitutions/bp, 95% CI [0.051–0.063]), and then intergenic sequence (0.046 substitutions/bp, 95% CI [0.044–0.051]). These three sequence classes are significantly different from each other under pairwise bootstrapping tests (fourfold synonymous vs. introns: $P = 0.018$; introns vs. intergenic: $P < 0.001$; fourfold synonymous vs. intergenic: $P < 0.001$). The proximal and deep subdivisions

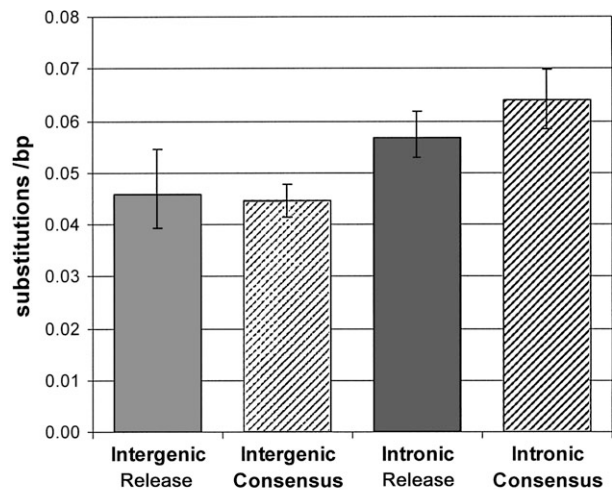


FIG. 2.—Raw divergence estimates for intergenic and intronic sites in the *Plasmodium falciparum* and *Plasmodium reichenowi* genomes as classified by the official release annotation and an annotation based on a strict consensus between the release annotation and automated gene predictions by the Glimmer M, Genefinder, and Phat software programs. Error bars indicate 95% CI determined through nonparametric bootstrapping process with 1,000 replicates. For both the intergenic and intronic sequence classes there was no significant difference between the divergence estimates obtained from sites identified under the release annotation and sites identified under the more stringent strict consensus annotation.

of the intergenic sequence class were not significantly different from each other ($P = 0.122$), with respective divergence levels of 0.048 substitutions/bp (95% CI [0.044–0.051]) and 0.042 substitutions/bp (95% CI [0.036–0.051]). Rich et al. (1998) reported a divergence level of 0.0378 substitutions/bp under a Jukes-Cantor model for fourfold synonymous sites in the *Rap1* gene of *P. falciparum* and *P. reichenowi*, which is lower than our divergence estimate for any sequence class in these genomes. This difference in divergence estimates may be due to an unusually low mutation rate in the region of *Rap1*, or, more likely, base-calling errors in the short, low-sequencing-coverage *P. reichenowi* contigs. Base-calling errors may inflate our divergence estimates for each sequence class above their absolute level but are not likely to bias the relative rates of divergence of the sequence classes.

To address the possibility that conserved coding sequences overlooked in the release annotation might depress the divergence rate of the intergenic and/or intronic sequence classes relative to the coding sequence class, we recalculated the divergence estimates using an annotation based on a strict consensus between the release annotation and three automated annotations. The strict consensus data sets were much smaller than the release data sets, with 759,050 bp of alignable intergenic sequence and 121,157 bp of alignable intronic sequence. Figure 2 displays the divergence levels of the intergenic and intronic sequence classes under the release annotation and strict consensus annotation. The strict consensus divergence estimates for the intergenic and intronic sequence classes are, respectively, 0.045 substitutions/bp (95% CI [0.042–0.048]) and 0.064 substitutions/bp (95% CI [0.057–0.072]). These two sequence classes remain significantly

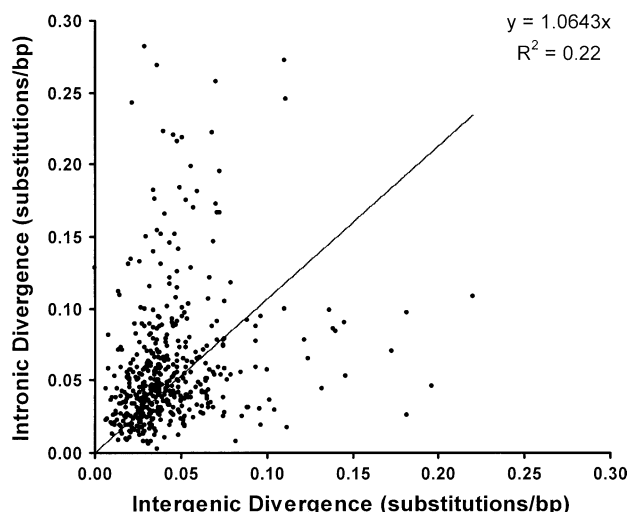


FIG. 3.—Scatterplot and linear regression of intergenic and intronic divergence estimates derived from 481 *Plasmodium reichenowi* genomic contigs containing at least 50 alignable base pairs of each sequence class. R^2 value indicates a weak but statistically significant correlation ($P < 0.001$).

different ($P < 0.001$) under the strict consensus annotation. Neither strict consensus divergence estimate differs significantly from the equivalent release estimate (intergenic: $P = 0.307$; intronic: $P = 0.077$). The higher divergence estimate for the intronic sequence under the strict consensus annotation, which is nearly statistically significant, may indicate that some sequences labeled as introns under the release annotation are actually conserved coding sequences.

We also tested for the possibility that the greater divergence level of intronic sequence relative to intergenic sequence may be caused by strong spatial variation in the background genomic mutation rate rather than an inherent difference in selective constraint between the sequence classes. Spatial heterogeneity in the background mutation rate has been detected in the human genome (Webster et al. 2004) and *Drosophila* genome (Singh, Arndt, and Petrov 2004). If gene-dense regions of the malaria genome experience a higher mutation rate than gene-poor regions, then introns may exhibit a higher divergence level simply by virtue of their greater inherent proximity to genes. In this case, paired divergence values of the intergenic and intronic sequence classes from the same genome fragments should be correlated. We performed a linear regression analysis of intronic and intergenic divergence estimates from 481 genome fragments that contained at least 50 bp of both intergenic and intronic sequences (fig. 3). The correlation between divergence rates is small but statistically significant ($R^2 = 0.22$; $P < 0.001$). This weak correlation is not likely to affect our conclusions, however, as the difference in the divergence levels of intergenic and intronic sequences from this subset of 481 genomic fragments is even greater than that in the overall comparison (intergenic divergence = 0.035 substitutions/bp, 95% CI [0.031–0.038], 482,740 bp total; intronic divergence = 0.054 substitutions/bp, 95% CI [0.048–0.060], 189,909 bp total). Though the significant correlation suggests variation in the background mutation rate in these genomes, the magnitude of the var-

iation is insufficient to mask an inherent difference in the rate of divergence of intergenic and intronic sequences.

Subramanian and Kumar (2003) observed a higher substitution rate in synonymous coding sites relative to noncoding sites in primate genomes, which they attributed to a greater incidence of hypermutable CpG dinucleotides in the coding sequence. When methylated, cytosine residues in CpG dinucleotides can undergo transitions to thymine at a rate approximately an order of magnitude faster than the normal rate of point substitution (Kondrashov 2003). The genome of *P. falciparum* is known to be at least partially methylated as inferred from restriction digest assay (Pollack, Kogan, and Golenser 1991), suggesting that some sequence classes may be susceptible to elevated mutation rates at CpG dinucleotides.

The overall GC content, CpG incidence, and observed/expected ratio of CpG dinucleotides for the various sequence classes in *P. falciparum* are presented in figure 4. GC content is largely similar among the sequence classes, with only a slight elevation at fourfold synonymous sites relative to noncoding sites. The incidence of CpG dinucleotides is sixfold higher in fourfold synonymous sites relative to the other sequence classes, however, due to the higher GC content at codon positions 1 and 2. Despite the increased incidence of CpG dinucleotides at fourfold sites, the observed number of CpGs is significantly lower than expected (2,183 observed vs. 2,983 expected; $P < 0.001$) given the frequencies of guanine and cytosine residues in codon positions 1, 2, and 3. This indicates that coding CpG dinucleotides are likely methylated and subject to a high mutation pressure. The approximate expected numbers of CpG dinucleotides are observed in introns (1,225 observed vs. 1,211 expected; $P = 0.313$) and intergenic sites greater than 500 bp from the coding sequence (2,309 observed vs. 2,301 expected; $P = 0.133$), indicating that these regions are likely not methylated. Surprisingly, there is an excess of CpG dinucleotides in the proximal intergenic sequence class (4,501 observed vs. 4,122 expected; $P < 0.001$), which may indicate the possible importance of CpG dinucleotides in gene regulation as well as a sharply demarcated boundary between methylated coding sequences and their flanking regions.

Divergence estimates for the sequence classes corrected for CpG mutations are presented in figure 1*b*. As one might predict, the CpG-rich fourfold synonymous site's sequence class was no longer significantly more divergent than the intronic sequence class (0.048 vs. 0.055 changes/bp; $P = 0.074$) or the intergenic sequence class (0.044 changes/bp; $P = 0.135$). However, the intronic sequence class remains significantly more divergent than the intergenic class after correction for CpG mutations ($P = 0.002$). Barring an intron-specific mutational process may indicate a greater degree of selective constraint on the intergenic sequence relative to introns in the *P. falciparum* and *P. reichenowi* genomes.

Conclusion

Transcribed fourfold synonymous coding sites exhibit the greatest divergence level of any sequence class in these genomes, though this elevated rate can be accounted for by

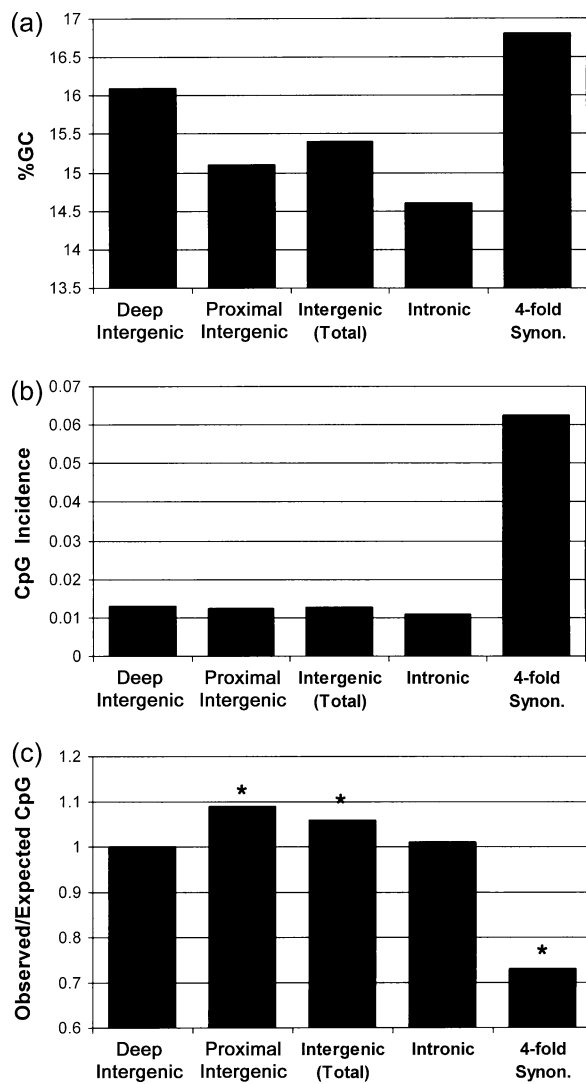


FIG. 4.—(a) GC content of *Plasmodium falciparum* noncoding sequence classes and fourfold synonymous coding sites that are alignable to orthologous regions in the *Plasmodium reichenowi* genome. (b) Incidence of CpG dinucleotides sites in aligned *P. falciparum* and *P. reichenowi* noncoding and fourfold synonymous coding regions. Sites were counted as CpG dinucleotides if one or both genomes exhibited a CpG dinucleotide in a given aligned position. (c) The ratio of observed/expected incidence of CpG dinucleotides in alignable noncoding and fourfold synonymous coding regions in *P. falciparum* and *P. reichenowi*. Asterisks denote significant deviations from the expectation as determined by a chi-square test (proximal intergenic: $P < 0.001$; intergenic [total]: $P < 0.001$; fourfold synonymous: $P < 0.001$).

accelerated substitutions at hypermutable CpG dinucleotides. This result provides strong corroboration of restriction digest assay, data that indicate the occurrence of methylation in the *P. falciparum* genome (Pollack, Kogan, and Golenser 1991). Even after a correction for accelerated divergence at CpG sites is taken into account, intronic sites are more divergent than intergenic sites. The reason for this pattern is unclear but may be related to the distribution of regulatory sites in these genomes. The surplus of intergenic CpG dinucleotides within 500 bp of the coding sequences indicates that conserved regulatory sites may be abundant in

this class of the sequence. Little is known regarding gene regulation in *Plasmodium*, but one recently characterized regulatory motif unique to *Plasmodium*, the G box, is overrepresented in the 5' flanking sequence of coding regions and can contain one or more CpG dinucleotides (Militello et al. 2004).

Interspecific divergence patterns offer great power to aid the interpretation of intraspecific polymorphism levels. It is a fundamental tenet of population genetics that divergence rates between species are proportional to average polymorphism levels within species under equilibrium conditions. Differences between sequence classes in the ratio of polymorphism to divergence are therefore indicative of recent selective or demographic events. In a comparison of the *P. falciparum* and *P. reichenowi* mitochondrial genomes, Conway et al. (2000) detected an eightfold higher level of divergence at synonymous coding sites relative to intergenic sites (0.1201 changes/bp vs. 0.0147 changes/bp) but approximately equal levels of polymorphism in those two sequence classes in six *P. falciparum* strains. To date there are no genome-wide estimates of polymorphism levels in different classes of nuclear sites in *P. falciparum* with which to compare the present analysis of interspecific divergence at nuclear sites. Our observation of substantial divergence across all classes of noncoding and silent sites appears to be discordant with the virtual absence of polymorphism at these sites, however, and together with the patterns of mitochondrial divergence observed by Conway et al. (2000) suggests that the contemporary rarity of polymorphisms in the *P. falciparum* genome may be a transient phenomenon, perhaps resulting from a recent population bottleneck.

Acknowledgments

The *P. reichenowi* sequence data were produced by the Pathogen Sequencing Group at the Wellcome Trust Sanger Institute and can be obtained from ftp://ftp.sanger.ac.uk/pub/pathogens/P_reichenowi/. This work was supported in part by National Institutes of Health grant GM 61351 and by a training grant in Global Infectious Diseases from the Ellison Medical Foundation.

Literature Cited

- Altschul, S. F., W. Gish, W. Miller, E. W. Myers, and D. J. Lipman. 1990. Basic local alignment search tool. *J. Mol. Biol.* **215**:403–410.
- Arnot, D. E. 1991. Possible mechanisms for the maintenance of polymorphisms in *Plasmodium* populations. *Acta Leiden.* **60**:29–35.
- Ayala, F. J., A. A. Escalante, and S. M. Rich. 1999. Evolution of *Plasmodium* and the recent origin of the world populations of *Plasmodium falciparum*. *Parasitol.* **41**:55–68.
- Cawley, S. E., A. I. Wirth, and T. P. Speed. 2001. Phat—a gene finding program for *Plasmodium falciparum*. *Mol. Biochem. Parasitol.* **118**:167–174.
- Conway, D. J., C. Fanello, J. M. Lloyd et al. (12 co-authors). 2000. Origin of *Plasmodium falciparum* malaria is traced by mitochondrial DNA. *Mol. Biochem. Parasitol.* **111**:163–171.
- Delcher, A. L., D. Harmon, S. Kasif, O. White, and S. L. Salzberg. 1999. Improved microbial gene identification with GLIMMER. *Nucleic Acids Res.* **27**:4636–4641.

- Escalante, A. A., and F. J. Ayala. 1994. Phylogeny of the malarial genus *Plasmodium*, derived from rRNA gene sequences. *Proc. Natl. Acad. Sci. USA* **91**:11373–11377.
- Forsdyke, D. R. 2002. Selective pressures that decrease synonymous mutations in *Plasmodium falciparum*. *Trends Parasitol.* **18**:411–417.
- Gardner, M. J., N. Hall, E. Fung et al. (45 co-authors). 2002. Genome sequence of the human malaria parasite *Plasmodium falciparum*. *Nature* **419**:498–511.
- Golightly, L. M., W. Mbacham, J. Daily, and D. F. Wirth. 2000. 3' UTR elements enhance expression of Pgs28, an ookinete protein of *Plasmodium gallinaceum*. *Mol. Biochem. Parasitol.* **105**:61–70.
- Hartl, D. L. 2004. The origin of malaria: mixed messages from genetic diversity. *Nat. Rev. Microbiol.* **2**:15–22.
- Hughes, A. L. 1999. Adaptive evolution of genes and genomes. Oxford University Press, Oxford.
- Hughes, A. L., and F. Verra. 1998. Ancient polymorphism and the hypothesis of a recent bottleneck in the malaria parasite *Plasmodium falciparum*. *Genetics* **150**:511–513.
- . 2001. Very large long-term effective population size in the virulent human malaria parasite *Plasmodium falciparum*. *Proc. R. Soc. Lond. B Biol. Sci.* **268**:1855–1860.
- . 2002. Extensive polymorphism and ancient origin of *Plasmodium falciparum*. *Trends Parasitol.* **18**:348–351.
- Jongwutiwes, S., C. Putaporntip, R. Friedman, and A. L. Hughes. 2002. The extent of nucleotide polymorphism is highly variable across a 3-kb region on *Plasmodium falciparum* chromosome 2. *Mol. Biol. Evol.* **19**:1585–1590.
- Kondrashov, A. S. 2003. Direct estimates of human per nucleotide mutation rates at 20 loci causing Mendelian diseases. *Hum. Mutat.* **21**:12–27.
- Militello, K. T., M. Dodge, L. Bethke, and D. F. Wirth. 2004. Identification of regulatory elements in the *Plasmodium falciparum* genome. *Mol. Biochem. Parasitol.* **134**:75–88.
- Pizzi, E., and C. Frontali. 1999. Molecular evolution of coding and non-coding regions in *Plasmodium*. *Parassitologia* **41**:89–91.
- . 2000. Divergence of noncoding sequences and of insertions encoding nonglobular domains at a genomic region well conserved in plasmodia. *J. Mol. Evol.* **50**:474–480.
- Pollack, Y., N. Kogan, and J. Golenser. 1991. *Plasmodium falciparum*: evidence for a DNA methylation pattern. *Exp. Parasitol.* **72**:339–344.
- Polley, S. D., and D. J. Conway. 2001. Strong diversifying selection on domains of the *Plasmodium falciparum* apical membrane antigen 1 gene. *Genetics* **158**:1505–1512.
- Rich, S. M., and F. J. Ayala. 1998. The recent origin of allelic variation in antigenic determinants of *Plasmodium falciparum*. *Genetics* **150**:515–517.
- . 2000. Population structure and recent evolution of *Plasmodium falciparum*. *Proc. Natl. Acad. Sci. USA* **97**:6994–7001.
- Rich, S. M., M. C. Licht, R. R. Hudson, and F. J. Ayala. 1998. Malaria's Eve: evidence of a recent population bottleneck throughout the world populations of *Plasmodium falciparum*. *Proc. Natl. Acad. Sci. USA* **95**:4425–4430.
- Rich, S. M., M. U. Ferreira, and F. J. Ayala. 2000. The origin of antigenic diversity in *Plasmodium falciparum*. *Parasitol. Today* **16**:390–396.
- Saul, A. 1999. Circumsporozoite polymorphisms, silent mutations and the evolution of *Plasmodium falciparum*. *Parasitol. Today* **15**:38–40.
- Saul, A., and D. Battistutta. 1988. Codon usage in *Plasmodium falciparum*. *Mol. Biochem. Parasitol.* **27**:35–42.
- Singh, N. D., P. F. Arndt, and D. A. Petrov. 2004. Genomic heterogeneity of background substitutional patterns in *Drosophila melanogaster*. *Genetics* **169**:709–722.
- Subramanian, S., and S. Kumar. 2003. Neutral substitutions occur at a faster rate in exons than in noncoding DNA in primate genomes. *Genome Res.* **13**:838–844.
- Thompson, J. D., D. G. Higgins, and T. J. Gibson. 1994. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.* **22**:4673–4680.
- Verra, F., and A. L. Hughes. 2000. Evidence for ancient balanced polymorphism at the apical membrane antigen-1 (AMA-1) locus of *Plasmodium falciparum*. *Mol. Biochem. Parasitol.* **105**:149–153.
- Volkman, S. K., A. E. Barry, E. J. Lyons, K. M. Nielsen, S. M. Thomas, M. Choi, S. S. Thakore, K. P. Day, D. F. Wirth, and D. L. Hartl. 2001. Recent origin of *Plasmodium falciparum* from a single progenitor. *Science* **293**:482–484.
- Volkman, S. K., D. L. Hartl, D. F. Wirth, et al. (11 co-authors). 2002. Excess polymorphisms in genes for membrane proteins in *Plasmodium falciparum*. *Science* **298**:216–218.
- Watanabe, J., M. Sasaki, Y. Suzuki, and S. Sugano. 2002. Analysis of transcriptomes of human malaria parasite *Plasmodium falciparum* using full-length enriched library: identification of novel genes and diverse transcription start sites of messenger RNAs. *Gene* **291**:105–113.
- Webster, M. T., N. G. Smith, M. J. Lercher, and H. Ellegren. 2004. Gene expression, synteny, and local similarity in human non-coding mutation rates. *Mol. Biol. Evol.* **21**:1820–1830.

Robin Bush, Associate Editor

Accepted April 14, 2005