Contrasting Histories of G6PD Molecular Evolution and Malarial Resistance in Humans and Chimpanzees

Brian C. Verrelli,*† Sarah A. Tishkoff,*‡ Anne C. Stone,*§ and Jeffrey W. Touchman*†‡¶

*Center for Evolutionary Functional Genomics, The Biodesign Institute, Tempe, Arizona; ‡School of Life Sciences, Arizona State University, Tempe; †Department of Biology, University of Maryland, College Park; §School of Human Evolution and Social Change, Arizona State University, Tempe; and ¶The Translational Genomics Research Institute, Phoenix, Arizona

Although mutations in the glucose-6-phosphate dehydrogenase (G6PD) gene result in several blood-related diseases in humans, they also confer resistance to malarial infection. This association between G6PD and malaria was supported by population genetic analyses of the G6PD locus, which indicated that these mutations may have recently risen in frequency in certain geographic regions as a result of positive selection. Here we characterize nucleotide sequence variation in a 5.2-kb region of the G6PD locus in a population sample of 56 chimpanzees, as well as among 7 other nonhuman primates, to compare with that in humans in determining whether other primates that are impacted by malaria also exhibit patterns of G6PD polymorphism or divergence consistent with positive selection. We find that chimpanzees have several amino acid variants but that the overall pattern at G6PD in chimpanzees, as well as in Old and New World primates in general, can be explained by recent purifying selection as well as strong functional constraint dating back to at least 30–40 MYA. These comparative analyses suggest that the recent signature of positive selection at G6PD in humans is unique.

Introduction

The combination of molecular and functional evolutionary genetics can be a powerful approach in determining how single mutations explain various human diseases. One of the best examples is that of the glucose-6-phosphate dehydrogenase (G6PD [MIM 305900]) enzyme where functional analyses have shown that naturally occurring mutations within this gene reduce enzyme activity, some to the degree that they result in various cardiovascular disorders (Luzzatto et al. 2001; Beutler and Vulliamy 2002). Interestingly, these G6PD enzyme–deficiency mutations are frequent in specific geographic regions and are associated with resistance to malarial infection by Plasmodium parasites (Vulliamy et al. 1992; Ruwende et al. 1995). Because malaria is one of the leading causes of death, affecting more than 400 million people worldwide (Greenwood and Mutabingwa 2002), G6PD is the target of much research to understand the history of resistance to malaria (Tishkoff and Verrelli 2004).

Although G6PD variants may be adaptive, until recent molecular genetic analyses, this hypothesis was unsubstantiated. Population genetic analyses from human populations inside and outside of Africa found strong linkage disequilibrium (LD) only associated with amino acid replacement-bearing haplotypes, which supported a recent increase in frequency for these variants in the last 20,000 years (Tishkoff et al. 2001; Sabeti, Reich et al. 2002; Saunders et al. 2002, 2005), consistent with the onset of severe malaria (Lingstone 1971). In addition, relative to a neutral model of molecular evolution, which predicts that variation accumulates as a result of genetic drift, comparative species and coalescent analyses show that G6PD variants are higher in number and frequency than expected (Verrelli et al. 2002; Coop and Griffiths 2004). These analyses suggest that G6PD variants have been recently maintained under selective pressure from malaria and played a large role in disease resistance.

Although humans serve as a host for several Plasmodium species, it is not clear how severe malaria has been in other primates (Ollomo et al. 1997; Sullivan et al. 2001; Martin et al. 2005). In primates, many Plasmodium species are also only associated with specific taxa, possibly having coevolved with them (Escalante et al. 2005). For example, serious blood-related complications and death in humans are largely the result of infection from Plasmodium falciparum, whereas less severe symptoms are also associated with Plasmodium vivax, Plasmodium malariae, and Plasmodium ovale infection (Miller et al. 2002; Rasti et al. 2004). The Plasmodium fieldi, Plasmodium inui, and Plasmodium knowlesi strains are associated with nonhuman primate malaria in Southeast Asia in macaques, and Plasmodium reichenowi has been isolated from both chimpanzees (Pan troglodytes) and gorillas (Gorilla gorilla) in their natural habitats in Africa (see Coatey et al. 1971; Ollomo et al. 1997). Nonetheless, these studies find that severe malarial symptoms are not always associated with Plasmodium infection in nonhuman primates to the degree that we observe in humans.

One approach to determining whether malaria has historically impacted other closely related primates and has solicited a genetic response for resistance is through comparative species analyses of genes like G6PD, which we find to exhibit adaptive signatures of selection within human populations (e.g., Kwiatkowski 2005). The chimpanzee, P. troglodytes, is our closest relative having diverged from the human lineage approximately 5–7 MYA (e.g., Kumar et al. 2005). Relative to human genetic analyses, there is still very little information about disease resistance and population genetic variation in general in chimpanzees, and yet their genome differs from humans by only approximately 2% (Cheng et al. 2005; Chimpanzee Sequencing and Analysis Consortium 2005). In this respect, given that epidemiological studies are often difficult to conduct with nonhuman primates, especially within their natural habitats, comparative molecular approaches among species with highly similar genomes can be powerful in elucidating the functional and evolutionary history of disease resistance.

Key words: glucose-6-phosphate dehydrogenase, malarial resistance, chimpanzee.

E-mail: brian.verrelli@asu.edu.

doi:10.1093/molbev/msi024
Advance Access publication June 2, 2006

© The Author 2006. Published by Oxford University Press on behalf of the Society for Molecular Biology and Evolution. All rights reserved.
For permissions, please e-mail: journals.permissions@oxfordjournals.org
The goal of this current study is to characterize the nucleotide sequence variation at the G6PD locus in a sample of closely related nonhuman primates to determine how the pattern of adaptive amino acid variation found in humans compares with the evolutionary history of primates in general. Here we collect data from a population sample of chimpanzees from various subspecies, in addition to 7 other nonhuman primates, for comparison with a global sample of humans. Specifically, we are interested in whether the signatures of recent positive selection found in humans, such as the excess of G6PD amino acid variation or the long-range pattern of LD, reflect a general pattern of molecular evolution at G6PD in primates or whether it is a signature that is unique to the human lineage.

Materials and Methods

Primate Samples

Nucleotide sequences were collected from DNA samples of 46 unrelated chimpanzees: 39 from the West African Pan troglodytes verus, 5 from the Central African Pan troglodytes troglodytes, 1 from the East African Pan troglodytes schweinfurthii, and 1 Nigerian Pan troglodytes vellerosus, which has been purported to be an additional subspecies (Gonder et al. 1997). Samples were imported in accordance with the Convention on International Trade in Endangered Species of Wild Fauna and Flora under permit 99US013176/9. DNA samples of 7 primates were also obtained from the Coriell Institute for Medical Research (http://www.coriell.org/), including Old World apes: bonobo (Pan paniscus), gorilla (Gorilla gorilla gorilla), orangutan (Pongo pygmaeus), siamang (Hylobates syndactylus), macaque (Macaca mulatta), and baboon (Papio anubis), and a New World monkey, the tamarin (Saguinus oedipus). In addition to these new data, from Verrelli et al. (2002), we add G6PD sequences of 216 human, 3 P. t. verus, and 1 P. t. troglodytes for a total of 42 P. t. verus; however, each of these 6 unrelated individuals had 2 different haplotypes that were resolved empirically (see below), resulting in a total of 48 P. t. verus sequences.

Polymerase Chain Reaction and Sequence Determination

All polymerase chain reaction (PCR) primers used in this study are based on the GenBank accession number X55448 and are available on request. Because we are primarily interested in comparing variation at the G6PD locus within and between human and chimpanzee populations, we PCR amplified the same 5.2-kb region from the study of Verrelli et al. (2002) that spans exons 3–13 (fig. 1), which contain the majority of human G6PD-deficiency mutations recorded to date (Beutler and Vulliamy 2002). Exon 1 is not translated, and exon 2 is separated from exon 3 by the approximately 10-kb intron 2, which includes many Alu repeats. The 5.2-kb region was analyzed for all samples except for the macaque, baboon, and tamarin, for which only coding regions were amplified and sequenced for interspecific comparisons. PCR products were prepared using shrimp alkaline phosphatase and exonuclease I (US Biochemicals, Cleveland, OH) and sequenced with ABI Big Dye terminator kit and Prism 377 DNA Sequencer (Applied Biosystems, Foster City, CA). Sequence files were aligned using the Sequencher v. 4.5 software program (Gene Codes, Ann Arbor, MI).

For the bonobo, gorilla, orangutan, siamang, macaque, baboon, and tamarin samples, only a single G6PD sequence was analyzed. These taxa served as outgroup samples here, and therefore, all heterozygous sites were ignored in analyses. In the case of the aforementioned 6 P. t. verus females, the 5.2-kb PCR product was cloned using the Invitrogen TOPO-XL kit, and a single clone was sequenced along with the initial PCR product. All single nucleotide polymorphisms (SNPs) identified in the clone haplotype sequence that did not appear as heterozygous in the initial PCR sequence were considered cloned PCR artifacts and ignored.

Data Analysis

Our chimpanzee individuals reflect a random sample in that they were not chosen based on phenotype or specific G6PD allelic variation. Therefore, population genetic parameter estimates and statistical tests of neutrality for our sample should reflect the frequencies of variants found in the natural population. Humans from sub-Saharan Africa...
typically exhibit significantly greater genetic diversity compared with other regions, which is consistent with a larger and more ancestral effective population size ($N_e$) for sub-Saharan groups and a relatively recent ancestry and smaller $N_e$ for other regions (Tishkoff and Williams 2002; Tishkoff and Verrelli 2003). These differences in demographic histories can result in distinct patterns of LD and haplotype structure; therefore, as in the Verrelli et al. (2002) analysis, we present estimates of diversity from our African and non-African humans separately, in addition to estimates from the global human population. As with humans, we present population genetic estimates for the overall *P. troglodytes* species group as well as comparisons among the subspecies.

For all of our population and comparative species genetic analyses, unless otherwise stated, we have used the Rozas et al. (2003) DnaSP v. 4.0 program. Silent SNPs (which refers to introns and synonymous sites in exons throughout the study) do not alter the amino acid sequence; therefore, they are expected to best reflect neutral processes. Locus-specific estimates of diversity based on the number of segregating sites ($S$), and corrected by sample size, were calculated using Watterson’s $\theta$ (1975). These estimates were compared with $\pi$, which is an estimate based on the average number of pairwise differences among sequences. These 2 estimates of the parameter $3N_e\mu$ (for X-linked genes, where $\mu$ is the mutation rate) were calculated for chimpanzees and humans for silent and replacement sites. The 2 estimates, $S$ and $\pi$, are expected to be equal under neutrality, which can be assessed using Tajima’s (1989) test. Positive $D$ values may be consistent with balancing selection whereas negative $D$ values suggest directional selection. Changes in demographic history, such as population structure or expansion, may also be consistent with positive or negative $D$ values, respectively. We can compare these $D$ values with human *G6PD* as well as with other chimpanzee loci (i.e., Yu et al. 2003; Fischer et al. 2004) to determine how selection and demography have played roles in shaping *G6PD* SNP distributions in the 2 species.

Under a model of neutrality, we expect that polymorphism and fixation are inherently correlated, and thus, the relative amount of between-species fixation should reflect the amount of within-species polymorphism (McDonald and Kreitman 1991). We compared the level of chimpanzee *G6PD* polymorphism at silent and replacement sites with the level of fixation between chimpanzee and human *G6PD* at silent and replacement sites with the McDonald–Kreitman (1991) test. In addition, we also examined the historical functional constraint acting at *G6PD* by contrasting levels of amino acid fixation among our 9 primate species samples. We conducted pairwise tests of the ratio of nonsynonymous to synonymous site fixation ($d_{s}/d_{s}$) in coding regions among the 9 samples, where a $d_{s}/d_{s} > 1$ is typically interpreted as evidence for positive selection and $d_{s}/d_{s} < 1$ as evidence for purifying selection on amino acid variation.

Finally, we were interested in how the haplotype structure and LD at *G6PD* in chimpanzees compare with those found at human *G6PD*. We used the LDhat program of McVean et al. (2002), which applies the approximate-likelihood method of Hudson (2001), to estimate the population parameter $\rho = 3N_e\lambda$ (where $\lambda$ is the recombination rate per bp) across the 5.2-kb region. The LDhat program then conducts a permutation analysis to determine whether pairwise comparisons among SNPs exhibit significantly more or less LD, given the SNP allele frequencies and the previously generated gene-specific estimate of $\rho$. Tests of association involving rare SNPs, especially singletons, will not have the statistical power to result in significance. Therefore, as determined by the LDhat program, only SNPs that were informative for estimates of recombination were used in analyses of LD in the human and chimpanzee population samples.

### Results

**G6PD Chimpanzee Population Variation**

There are 1425 bp in the exons 3–13 and 3725 bp in the noncoding regions, including introns and the 3’ untranslated region, for a total of 5150 bp of *G6PD* sequence (fig. 1). Excluding replacement sites in exons and conserved splice sites at the beginning and end of the introns, there are a total of 4017 effectively silent sites that are used for all estimates of silent site diversity at *G6PD*. All estimates

<table>
<thead>
<tr>
<th>Samplea</th>
<th>Silentb</th>
<th>Replacement</th>
<th>MKd</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$n^b$</td>
<td>$S^c$</td>
<td>$\pi^c$</td>
</tr>
<tr>
<td><em>Pan troglodytes</em></td>
<td>56</td>
<td>29</td>
<td>0.079</td>
</tr>
<tr>
<td><em>Pan troglodytes verus</em></td>
<td>48</td>
<td>12</td>
<td>0.058</td>
</tr>
<tr>
<td><em>Pan troglodytes troglodytes</em></td>
<td>6</td>
<td>11</td>
<td>0.120</td>
</tr>
<tr>
<td><em>Homo sapiens</em></td>
<td>216</td>
<td>23</td>
<td>0.073</td>
</tr>
<tr>
<td>African</td>
<td>160</td>
<td>22</td>
<td>0.072</td>
</tr>
<tr>
<td>Non-African</td>
<td>56</td>
<td>8</td>
<td>0.053</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Samplea</th>
<th>Silentb</th>
<th>Replacement</th>
<th>MKd</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$n^b$</td>
<td>$S^c$</td>
<td>$\pi^c$</td>
</tr>
<tr>
<td><em>Pan troglodytes</em></td>
<td>56</td>
<td>29</td>
<td>0.079</td>
</tr>
<tr>
<td><em>Pan troglodytes verus</em></td>
<td>48</td>
<td>12</td>
<td>0.058</td>
</tr>
<tr>
<td><em>Pan troglodytes troglodytes</em></td>
<td>6</td>
<td>11</td>
<td>0.120</td>
</tr>
<tr>
<td><em>Homo sapiens</em></td>
<td>216</td>
<td>23</td>
<td>0.073</td>
</tr>
<tr>
<td>African</td>
<td>160</td>
<td>22</td>
<td>0.072</td>
</tr>
<tr>
<td>Non-African</td>
<td>56</td>
<td>8</td>
<td>0.053</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Samplea</th>
<th>Silentb</th>
<th>Replacement</th>
<th>MKd</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$n^b$</td>
<td>$S^c$</td>
<td>$\pi^c$</td>
</tr>
<tr>
<td><em>Pan troglodytes</em></td>
<td>56</td>
<td>29</td>
<td>0.079</td>
</tr>
<tr>
<td><em>Pan troglodytes verus</em></td>
<td>48</td>
<td>12</td>
<td>0.058</td>
</tr>
<tr>
<td><em>Pan troglodytes troglodytes</em></td>
<td>6</td>
<td>11</td>
<td>0.120</td>
</tr>
<tr>
<td><em>Homo sapiens</em></td>
<td>216</td>
<td>23</td>
<td>0.073</td>
</tr>
<tr>
<td>African</td>
<td>160</td>
<td>22</td>
<td>0.072</td>
</tr>
<tr>
<td>Non-African</td>
<td>56</td>
<td>8</td>
<td>0.053</td>
</tr>
</tbody>
</table>

Table 1

Population Diversity Estimates and Tests of Neutrality at *G6PD*

---

a Samples are described in Materials and Methods.
b Number of chromosomes.
c Includes coding and noncoding regions (see Results).
d McDonald–Kreitman test (see Results).
e Number of SNPs.
f Average pairwise sequence differences (%).
g Tajima’s $D$ test ($* P < 0.05$).
In our sample of 56 *P. troglodytes* sequences, there are 34 polymorphisms, including intronic 8-bp insertion and 21-bp deletion variants, and 32 SNPs, 24 of which are noncoding. Of the 8 coding region SNPs, 3 alter the amino acid sequence and appear as singletons in our 48 *P. t. verus* sequences at amino acid residues 375 (Asp/Tyr), 425 (Gly/Ser), and 430 (Asn/Asp). Interestingly, haplotype 23 has both the Ser425 (SNP3448) and Tyr375 (SNP3598) variants (fig. 2), the latter of which occurs in the NADP-binding domain (fig. 1) that is coded by exon 10 and amino acids 372–413. The Asn to Asp430 (SNP3717) variant is adjacent to the splice site between intron 10 and exon 11 and predicts a more negative net protein charge.

When we account for *G6PD* being X-linked (multiplying by 4/3 because X chromosomes have a theoretical $N_e$ that is three-fourths that of autosomes), our estimates of silent site diversity ($\pi_f$) for *P. t. troglodytes* overall (0.11%), *P. t. verus* (0.08%), and *P. t. troglodytes* (0.16%) samples are similar to estimates found at other coding and noncoding loci for these species groups (Deinard and Kidd 1999; Kaessmann et al. 1999; Stone et al. 2002; Gilad et al. 2003; Yu et al. 2003; Fischer et al. 2004; Wooding et al. 2005). The few previous population genetic analyses of *P. t. troglodytes* and *P. t. verus* find that the latter has a similar level of nucleotide diversity compared with humans, whereas the former has considerably more. Because our estimates of *G6PD* silent site diversity are in general agreement with other comparisons between humans and chimpanzees, this suggests that *G6PD* silent site diversity is not unusual in the 2 species and can be used for analyses in testing neutrality for amino acid replacement variation. Our estimates of divergence based on the average number of silent differences between haplotypes show that the diversity among *P. t. troglodytes* haplotypes is comparable to that among subspecies. However, on average, the amount of *G6PD* diversity between any two *P. t. verus* haplotypes is half as large as that found among subspecies (tables 1 and 2).
Finally, as in humans, silent site diversity at G6PD is skewed toward rare frequencies in chimpanzees (i.e., negative Tajima’s $D$, table 1). Although this was statistically significant for the overall *P. troglodytes* sample, when the different subspecies are independently analyzed, this apparent “excess” of rare alleles disappears and chimpanzee subspecies diversity looks very similar to humans. This “rare allele” artifact is explained by the mixing of population samples with different evolutionary histories (e.g., Ptak and Przeworski 2002). That is, several of the SNPs, which are singletons in our overall sample, may be common within or even fixed between subspecies, and therefore, when only a single sequence is sampled from these groups (e.g., *P. t. schweinfurthii* and *P. t. vellerosus*), these SNPs appear very rare in frequency. At this time, there are only a few chromosomal regions of this size analyzed in *P. t. verus* samples; nonetheless, our estimates of the SNP frequency distribution at G6PD are not considerably different from these previous estimates (Yu et al. 2003; Fischer et al. 2004; Wooding et al. 2005).

Chimpanzee and human G6PD share equal magnitudes of replacement SNPs, yet those found in chimpanzees are significantly rare in frequency (table 1). Although this appears also to be true of the non-African sample of humans, the apparent excess of rare replacement variants is the result of a single allele sampled in Tunisia that carries 2 of the 4 amino acid changes, whereas, in general, these variants are found at relatively common frequencies in specific populations (Verrelli et al. 2002). It is possible that with larger samples from each of the subspecies we may find more amino acid variants. However, given our sample size for *P. t. verus*, we are unlikely to have missed variants that are relatively common in nature and would otherwise continue to sample rare mutations. To evaluate this scenario, we generated 10 000 simulated sets of 50 sequences using a binomial sampling process. With these simulations, we would expect to fail to sample a SNP with a frequency of at least 10% less than 5% of the time. Therefore, sample size, at least for that of *P. t. verus*, is unlikely to explain the difference in frequency distributions for G6PD amino acid variants between humans and chimpanzees.

### Estimates of Recombination

In examining the association among variants across G6PD, we are interested in how recombination may have shaped haplotype diversity within populations. Given the amount of differentiation among subspecies relative to that found within *P. t. verus* at G6PD, an LD statistical analysis of the combined *P. troglodytes* sample would be inappropriate and uninformative. Therefore, we focused on our population sample of 48 *P. t. verus* sequences for comparisons with our human sample. For human G6PD, 15 of 25 SNPs in the African sample and 3 of 12 SNPs in the non-African sample met the LDhat likelihood criteria for estimating $\rho$ and significant LD. For both human samples, our maximum likelihood (ML) estimate of $\rho$ was less than 0.5 (not significantly different from 0). The African sample had 22 associations, of the possible 105, and the non-African sample had 3 associations, of the possible 3, which were significant from the LDhat permutation analysis. Of the 22 significant associations in the African sample, a single pair of SNPs exhibited significantly less LD than expected, given the SNP frequencies and likelihood estimate of $\rho \approx 0$. This high overall LD reflects only a single recombinant found among the entire sample of 216 sequences.

The LDhat program was used to determine that of the 12 silent SNPs in the 48 *P. t. verus*, 10 were informative for estimating $\rho$ and significant LD, resulting in an ML estimate of $\rho \approx 72$. Using this likelihood estimate, the LDhat permutation analysis finds 19 associations, of the possible 45, that are significant. However, of these 19 associations, 13 were due to significantly less LD between pairs than expected, in spite of the high overall gene estimate of $\rho$. This haplotype diversity among SNPs reflects the history of recombination that has decayed LD across the G6PD locus in *P. t. verus*. In fact, the Hudson (1987) estimate of recombination (i.e., four-gamete test) finds at least 4 crossing-over events in only 48 *P. t. verus* sequences, which starkly contrasts with patterns of LD and recombination in either of our human samples.

### G6PD Interspecific Comparisons

We find similar estimates of $\pi$ for G6PD in both humans and *P. t. verus* (table 1) and an estimate of 1.1% for silent site divergence between the 2 samples (table 2), both of which are observations consistent with other nuclear loci comparisons between the 2 species. Again, although the sample sizes of humans and *P. t. verus* differ by approximately 4-fold, all estimates of diversity are weighted by sample size, and therefore, these estimates are comparable between the 2 samples. Compared with the *P. t. verus* silent site diversity, and divergence between human and chimpanzee, the number of *P. t. verus* replacement SNPs is significantly higher than that expected under neutrality (McDonald–Kreitman test; table 1), which was a pattern similarly found in the human population sample. However, although the number of G6PD amino acid variants is significantly high for humans and *P. t. verus*, as noted above,
Table 3  

**Divergence at G6PD Coding and Noncoding Regions among Primates**

<table>
<thead>
<tr>
<th>Species Comparison</th>
<th>(d_a^a)</th>
<th>(d_b^b)</th>
<th>(d_c^c)</th>
<th>(d_{ds}/d_s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Human</td>
<td>Chimpanzee</td>
<td>0</td>
<td>2.13</td>
<td>1.04</td>
</tr>
<tr>
<td></td>
<td>Bonobo</td>
<td>0</td>
<td>1.82</td>
<td>1.00</td>
</tr>
<tr>
<td></td>
<td>Gorilla</td>
<td>0</td>
<td>1.82</td>
<td>1.62</td>
</tr>
<tr>
<td></td>
<td>Orangutan</td>
<td>0.09</td>
<td>3.06</td>
<td>3.74</td>
</tr>
<tr>
<td></td>
<td>Siamang</td>
<td>0.28</td>
<td>6.58</td>
<td>4.63</td>
</tr>
<tr>
<td>Chimpanzee</td>
<td>Bonobo</td>
<td>0</td>
<td>0.30</td>
<td>0.14</td>
</tr>
<tr>
<td></td>
<td>Gorilla</td>
<td>0</td>
<td>2.75</td>
<td>1.65</td>
</tr>
<tr>
<td></td>
<td>Orangutan</td>
<td>0.09</td>
<td>4.00</td>
<td>3.74</td>
</tr>
<tr>
<td></td>
<td>Siamang</td>
<td>0.28</td>
<td>7.57</td>
<td>4.72</td>
</tr>
<tr>
<td>Bonobo</td>
<td>Gorilla</td>
<td>0</td>
<td>2.44</td>
<td>1.65</td>
</tr>
<tr>
<td></td>
<td>Orangutan</td>
<td>0.09</td>
<td>3.69</td>
<td>3.75</td>
</tr>
<tr>
<td></td>
<td>Siamang</td>
<td>0.28</td>
<td>7.24</td>
<td>4.67</td>
</tr>
<tr>
<td>Gorilla</td>
<td>Orangutan</td>
<td>0.09</td>
<td>3.68</td>
<td>3.68</td>
</tr>
<tr>
<td></td>
<td>Siamang</td>
<td>0.28</td>
<td>7.24</td>
<td>4.57</td>
</tr>
<tr>
<td>Orangutan</td>
<td>Siamang</td>
<td>0.37</td>
<td>4.64</td>
<td>4.04</td>
</tr>
</tbody>
</table>

*a* Nonsynonymous fixed differences (% total of 1092 sites).  
*b* Synonymous fixed differences (% total of 333 sites).  
*c* Intronic and noncoding fixed differences (% total of 3684 sites).

their frequency distributions indicate that they have been under different selection pressures.

Table 3 shows that among humans, chimpanzees, and other Old World apes, all \(d_{ds}/d_s\) are less than 1, with none of the comparisons greater than 0.1. There are 333 silent sites in coding regions, and therefore, divergence is also shown for noncoding regions to increase our sample size for estimates of the “neutral” fixation rate. Nonetheless, comparisons of replacement to silent divergence still approximate zero, given the few fixed replacements. We extended this analysis with coding region variation from 3 additional primates and found that the trend of little G6PD amino acid divergence continues back to the split between Old and New World primates (fig. 3). If we examine the rate of change within major and minor lineages, we find that none of the branches have an estimate of \(d_{ds}/d_s\) > 1, let alone one that is greater than 0.1. In fact, there are only 8 estimated replacement fixations found scattered among the sampled primates (fig. 3), none of which alter the net charge of the protein or occur in the G6P- or NADP-binding domains (fig. 1).

**Discussion**

Although humans and chimpanzees appear to have similar numbers of G6PD amino acid variants segregating within their populations, amino acid replacement SNPs appear to be very rare in frequency in the chimpanzee population. Estimates of LD associated with G6PD amino acid variants in humans imply a very recent increase in their frequency, whereas analyses of the haplotype structure at the G6PD locus in chimpanzees imply a history of several recombination events and very little overall LD. Here we discuss how these different patterns may be interpreted and what implications they may have for understanding the evolution of malarial resistance within humans and related primates.

**Chimpanzee G6PD Variation**

Estimates of silent site nucleotide diversity within chimpanzees and divergence found between humans and chimpanzees at G6PD are similar to other gene estimates (e.g., Ebersberger et al. 2002; Gilad et al. 2003, Fischer et al. 2004). Also consistent with these studies is that *P. t. verus* shows similar silent site diversity compared with humans, whereas *P. t. troglodytes* is twice as diverse. In addition, although we have only a few sequences from *P. t. troglodytes, P. t. schweinfurthii*, and *P. t. vellerosus*, these groups appear to be almost twice as differentiated from each other as are any two *P. t. verus*. This difference in diversity is likely due to an elevated historical \(N_e\) associated with these subspecies (Deinard and Kidd 1999; Kaessmann et al. 1999; Stone et al. 2002; Fischer et al. 2004; Won and Hey 2005) compared with *P. t. verus* and humans.

Our smaller samples of subspecies have been informative about the relative level of silent and replacement site divergence among humans and chimpanzee subspecies. If we had found little silent site divergence among chimpanzee subspecies, we may expect to find little amino acid diversity as well. However, although our population samples of *P. t. verus* and humans are similar with respect to silent site nucleotide diversity, it is clear that *P. t. verus* and other chimpanzee subspecies are considerably differentiated from each other. Given our sample size of *P. t. verus*, it is unlikely that further sampling would result in anything but more rare-frequency amino acid variants. However, it will be interesting to further sample the other more diverse subspecies with predicted greater \(N_e\) to determine if they also harbor G6PD amino acid variants.

**Purifying Versus Positive Selection**

One of the first signatures of positive selection identified at G6PD in humans was that strong LD was only associated with certain replacement-bearing sites, indicating their recent increase in frequency by selection and not by population expansion (Tishkoff et al. 2001; Sabetti, Reich et al. 2002; Saunders et al. 2002, 2005).
We know very little about background levels of recombination and LD in the chimpanzee genome, and furthermore, we do not necessarily expect these estimates to be similar in humans and chimpanzees (Myers et al. 2005; Ptak et al. 2005). Nonetheless, we may still expect that if positive selection had recently favored replacement variation that we may find, as in humans, that LD is strongly associated only with replacement-bearing haplotypes. However, we find no excess LD at G6PD in chimpanzees, and in fact, we find several examples of historical recombination unlike that found at human G6PD. Therefore, even though P. t. verus and humans have similar levels of G6PD nucleotide diversity, patterns of LD and inferred recombination are considerably different. It is unclear whether these different patterns at G6PD for humans and chimpanzees are common throughout the genome. In this respect, a fine-scale estimate of recombination rates throughout the chimpanzee genome would expand our understanding of how diversity in human and chimpanzee genomes is generated.

Other than the well-studied Major Histocompatibility Complex loci (i.e., Gyllensten et al. 1990; Adams et al. 2000; de Groot et al. 2000, 2005), the number of chimpanzee nuclear loci with population-level amino acid polymorphisms (i.e., not rare singletons from protein screens) is relatively small (e.g., Dufour et al. 2000; Sanghera et al. 2001; Soldevila et al. 2004). This suggests that our observation of 3 G6PD amino acid variants in P. t. verus may be unusual. Although our McDonald–Kreitman test implies an excess of G6PD replacement polymorphism in P. t. verus relative to silent site diversity, echoing the pattern at G6PD in humans (Verrelli et al. 2002), the pattern of the SNP frequency distributions is considerably different in the 2 species. Our Tajima’s D analysis indicates that amino acid polymorphisms in global samples of humans are not significantly rare in frequency, such as the common G6PD A and A- variants. On the other hand, although some of the replacement SNPs in our human sample were rare in the overall population, such as the G6PD Med and Kaiping variants, they are common in specific geographic regions (Ganczakowski et al. 1995; Kaeda et al. 1995; Beutler and Vulliamy 2002; Drozsiotou et al. 2004). Although historical population structure and recent adaptation in humans may have resulted in population-specific G6PD variants that are otherwise low in frequency, this is unlikely to be true for P. t. verus. The current geographic range for P. t. verus appears to be relatively small within Africa (Gagneux et al. 2001), and we find very little evidence for structure within P. t. verus for silent site diversity both from our data as well as from others (e.g., Wooding et al. 2005). This lack of structure for silent SNPs suggests that the pattern of rare G6PD replacement SNPs cannot be explained by population structure.

Amino acid polymorphism is prevalent in the mtDNA genomes of chimpanzees and humans (Nachman et al. 1996; Hasegawa et al. 1998), as well as in Drosophila and mice (Ballard and Kreitman 1994; Rand and Kann 1996). However, these SNPs segregate at very rare frequencies, suggesting that they are removed by weak purifying selection and are only found as singletons because they are slightly deleterious (Ohta 1992). Due to their common frequencies, G6PD amino acid variants in humans are generally accepted as being consistent with a model of spatially varying selection coefficients that enables amino acid polymorphisms to be maintained in the overall population. On the contrary, the pattern of variation in the P. t. verus sample implies that amino acid polymorphisms are slightly deleterious at G6PD and removed by weak purifying selection. From a functional perspective, the fact that both the rare chimpanzee replacement-bearing haplotypes have variants near or within the NADP-binding site is consistent with their deleterious nature. Furthermore, relative to the level of G6PD silent site divergence across primates, there is very little G6PD protein evolution, even as far back as the split between New and Old World apes, approximately 30–40 MYA. These analyses imply that the evolutionary history of G6PD in primates has been subjected to long periods of strong functional constraint.

Protein conservation has long been expected for “housekeeping” genes such as G6PD that belong to the glycolytic pathway (e.g., Lewontin and Hubby 1966). However, coincidentally, one of the more dramatic examples of adaptive protein evolution is at G6PD in the fruit fly species Drosophila melanogaster and Drosophila simulans (Eanes et al. 1993; Eanes, Kirchner, Yoon et al. 1996). In addition, the one common D. melanogaster G6PD replacement SNP that varies in frequency with latitude and is likely under balancing selection for enzyme activity variation (Oakeshott et al. 1983) is found in the highly conserved NADP-binding site (Eanes, Kirchner, Taub et al. 1996). Although certain human variants (e.g., the Med allele) are found within the G6P- and NADP-binding domains, which generally cause a drastic reduction in enzyme activity, several common variants are found outside of the conserved binding domains (fig. 1). Based on our comparative analyses across species, it appears that G6PD replacement variation is not tolerated in the primate lineage and is largely deleterious within chimpanzees. In spite of this, it appears that the pattern of G6PD replacement variation in humans is more like that in Drosophila than other primates, in that it not only is adaptive but also can be beneficial when it occurs within the most conserved domains of the enzyme.

Malarial Selection at G6PD

Given our limited knowledge about malaria in nonhuman primates, what do these analyses of G6PD tell us? If G6PD was the only gene targeted for malarial resistance, we may conclude that severe malaria has not had an impact during chimpanzee evolution given the lack of signatures of recent or long-term evolutionary change for G6PD in this primate lineage. However, we know of many other malarial resistance genes in humans alone (e.g., Hamblin et al. 2002; Sabeti, Usen et al. 2002; Wang et al. 2003), so this is unlikely to be a valid conclusion. It is possible that the adaptive response was in the form of transcriptional regulation (e.g., cis–trans variation); nonetheless, this still is in stark contrast to the pattern seen for amino acid variation in humans.

From our current analyses, it is clear that G6PD has not been under recent positive selection for amino acid variation in chimpanzees, for malarial resistance or otherwise. In fact, it appears that G6PD evolution in chimpanzees
reflects the long-term history of purifying selection across \textit{G6PD} in primates in general. Our analyses suggest that comparative population genetic analyses between humans and chimpanzees can be very informative for identifying different signatures of malarial selection and, thus, different targets of selection in the 2 species. This has recently been shown with a similar comparative analysis of the CCR5 promoter region, which is linked to immune response, that showed different selective pressures historically acting on humans and chimpanzees (Wooding et al. 2005). In our case, we may rule out \textit{G6PD} as a candidate for further functional analyses of malarial resistance in \textit{P. t. verus}, given the limited evidence it shows for adaptive evolution.

Conclusion

Interspecific analyses can shed light on the forces that have shaped our human species and made us historically unique compared with other primates (e.g., Clark et al. 2003; Bustamante et al. 2005). However, without our population analyses of human and chimpanzee \textit{G6PD}, we would not have realized that amino acid variation is abundant in humans and that our species has recently responded to malarial infection differently than our closest relative. Although genomic analyses reveal that chimpanzees and humans differ by approximately 2%, these analyses ignore the variation within species that reflect our recent history. In fact, given that many diseases in humans, such as malaria, may have arisen in the last 10,000 years, we may often need both interspecific and intraspecific analyses of humans and chimpanzees to understand recent adaptation in our species.

Acknowledgments

We thank Stephanie Rogers for assistance with the macaque, baboon, and tamarin exon sequence annotation; George Perry for assistance with chimpanzee DNA sample coordination; and John H. McDonald as well as 3 anonymous reviewers for comments on the manuscript. Chimpanzee samples were generously provided by H. Vasken Aposhian, the Detroit Zoological Institute (Detroit, MI), the Lincoln Park Zoo (Chicago, IL), the Primate Foundation of Arizona (Mesa, AZ), the Riverside Zoo (Scottsbluff, NE), the Southwest Foundation for Biomedical Research (San Antonio, TX), the Sunset Zoo (Manhattan, KS), the Welsh Mountain Zoo (North Wales, UK), the Yerkes National Primate Research Center (Atlanta, GA), and the New Iberia Research Center (Lafayette, LA), where sample collection and maintenance was supported by a National Institutes of Health—National Center for Research Resources grant U42-RR015087 to the University of Louisiana at Lafayette. This work was supported by funding to B.C.V. from the Center for Evolutionary Functional Genomics in The Biodesign Institute and the School of Life Sciences at the Arizona State University.

Literature Cited


John H. McDonald, Associate Editor

Accepted May 16, 2006