Rapidly Evolving Genes in Human. I. The Glycophorins and Their Possible Role in Evading Malaria Parasites

Hurng-Yi Wang,†‡ Hua Tang,‡ C.-K. James Shen,† and Chung-I Wu‡*

*Department of Biology, National Taiwan Normal University, Taipei, Taiwan; †Institute of Molecular Biology, Academia Sinica, Taipei, Taiwan; ‡Department of Ecology and Evolution, University of Chicago

In an attempt to identify all fast-evolving genes between human and other primates, we found three glycophorins, GPA, GPB, and GPE, to have the highest rate of nonsynonymous substitutions among the 280 genes surveyed. The K_a/K_s ratios are generally greater than 3 for GPA, GPB, and GPE in human, chimpanzee, and gorilla, indicating positive selection. The uniformly high substitution rate across loci can be explained by the frequent sequence exchanges among genes. GPA is the receptor for the binding ligand EBA-175 of the malaria parasite, *Plasmodium falciparum*. The levels of nonsynonymous divergence and polymorphism of EBA-175 are also the highest in the genome of *P. falciparum*. We hypothesize that GPA has been evolving rapidly to evade malaria parasites. Both the high rate of nonsynonymous substitutions and the frequent interlocus conversions may be means of evasion. The support for the evasion hypothesis is still indirect, but, unlike other hypotheses, it can be tested specifically and systematically.

Introduction

The influx of DNA sequences has permitted a systematic analysis of rapidly evolving genes between closely related species. Such genes are enriched with information about the action of positive Darwinian selection and may shed light on the process of genic adaptation (Kitano and Saitou 1999; Wyckoff, Wang, and Wu 2000; Fay, Wyckoff, and Wu 2001; Johnson et al. 2001; Enard et al. 2002; Fay, Wyckoff, and Wu 2002; Smith and Eyre-Walker 2002; Bamshad and Wooding 2003). We have thus initiated an attempt at characterizing each rapidly evolving gene among higher primates. Previous studies of rapidly evolving genes have revealed a trend among genes of male reproduction (Civetta and Singh 1998; Ting et al. 1998; Wyckoff, Wang, and Wu 2000; Swanson et al. 2001) and defense against pathogens (Yang and Bielawski 2000). A systematic survey can confirm the generality of those observations and add interesting exceptional cases.

There are currently more than 500 published coding sequences from the Old World monkey (OWM) that can be inferred to be orthologous to a human gene. Among them, 280 meet a set of criteria (length, completeness, etc.; see *Materials and Methods*) to become the basis of our analysis. Of these 280 genes, 17 have been determined to be fast evolving. For each of these fast-evolving genes, we ask (1) if the gene has evolved especially rapidly in the human lineage; (2) if there are specific sites under positive selection; and (3) what may be the forces driving the rapid evolution.

In this data set, the fastest evolving genes in the human lineage are the glycophorins. There are three glycophorin loci in human, chimpanzee, and gorilla but only one in other primate species (Rearden et al. 1993; Blumenfeld et al. 1997). In human, glycophorin A (GPA) and B (GPB) code for antigens underlying the very common MN and Ss blood type polymorphisms, respectively. At least 40 other blood types are caused by the glycophorin variation (Blumenfeld and Huang 1995). Rearrangements by unequal recombination and/or gene conversion between glycophorin genes appear to be very common, and hot spots of recombination exist in a region of 4 kb encompassing the three extracellular exons (II, III, and IV) (Blumenfeld and Huang 1997).

While GPA constitutes the most abundant glycoproteins on the erythrocyte surface, an exceptional case of deletion homozygote for both GPA and GPB has been known to lead to normal adulthood (Schenkel-Brunner 2000). In human, GPA has been shown to be the receptor of a binding ligand, the 175-kD erythrocyte-binding antigen (EBA-175) of *Plasmodium falciparum* (Pasvol, Wainscoat, and Weatherall 1982; Sim et al. 1994). A recent study analyzed the evolution of GPA among higher primates (Baum, Ward, and Conway 2002) and suggested a "decoy" hypothesis for its rapid evolution. To further evaluate the possible forces driving the evolution of glycophorins, we sequenced the extracellular domain (exons II, III, and IV) of all three glycophorin genes from human, chimpanzee, and gorilla and the single gene from gibbon. We also analyzed published DNA sequences of *P. falciparum* in conjunction with the glycophorin sequences. An alternative "evasion" hypothesis is proposed to account for the overall patterns that are not easily discernible in the GPA sequences alone.

Materials and Methods

Data Collection

All DNA sequences of the OWM (mostly *Macaca* and *Papio*) were downloaded from GenBank. Coding region sequences (CDS) were trimmed from the GenBank records. For each OWM sequence, we chose as putative ortholog the human refseq that has the highest score and lowest *E* value in the blast search. Among more than 500 pairs of sequences so identified, we retained those with matching annotations between human and OWM and rejected those with many ambiguous codons. Finally, we calculated the K_a and K_s values for each putative pair of orthologs by the method of Li (1993), where K_a is the number of nonsynonymous substitutions per nonsynonymous site and K_s is the number of synonymous substitutions per synonymous site. We finalized the data set using the 280 genes with $K_s < 0.15$ (most with $K_s < 0.1$; the

Key words: positive selection, glycophorin, malaria, gene conversion, rapid evolution.

E-mail: ciwu@uchicago.edu.

Mol. Biol. Evol. 20(11):1795–1804. 2003
DOI: 10.1093/molbev/msg185
Molecular Biology and Evolution, Vol. 20, No. 11,

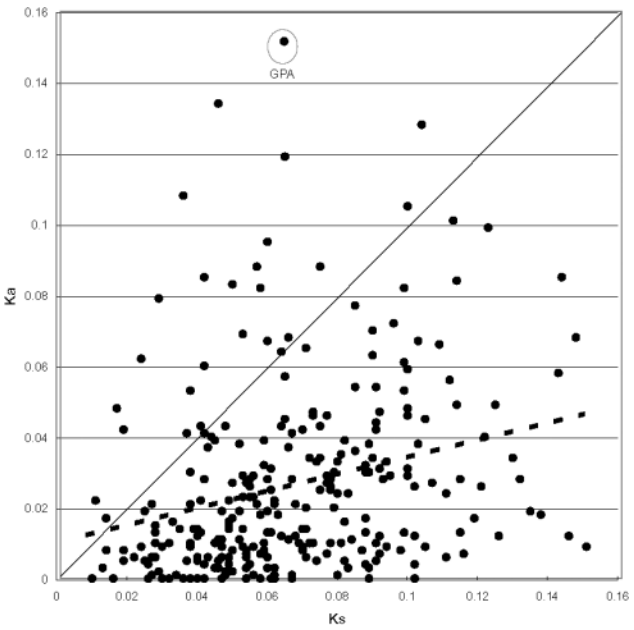

FIG. 1.—The K_a values of 280 orthologous genes between human and macaque are plotted against K_s. GPA is indicated (GPB and GPE are not shown).

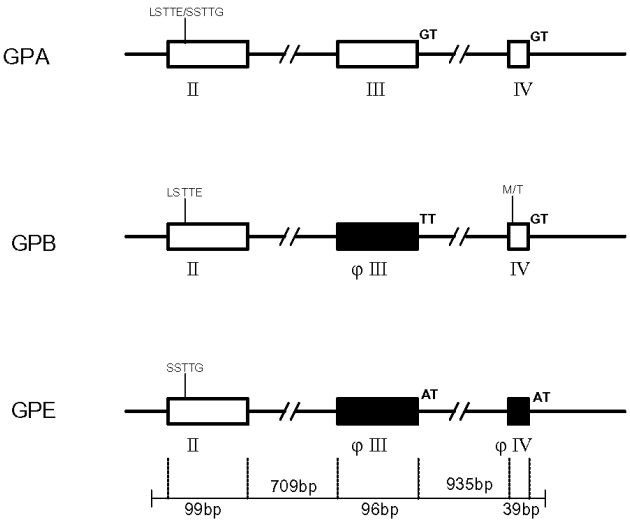

FIG. 2.—The canonical genomic structure of human GPA, GPB, and GPE. These genes are tandemly arranged and exons 2 to 4 account for 226 bp of the 2-kb region sequenced. The LSTTE/SSTTG stretches of amino acids represent the N/M antigenic determinants. The M/T amino acids in exon IV determine the S/s antigens. Unexpressed pseudoexons (φ) are represented by filled boxes followed by the splice site mutations.

average being 0.066). Because the objective is to compare K_a and K_s, the occasional misidentification of othology should not affect the main goal of searching for genes with high K_a/K_s ratios. Figure 1 contains these calculations.

DNA Sequencing

The extracellular domain of the glycophorin genes, which spans exons II, III, and IV (fig. 2), was targeted for polymerase chain reaction. Two primers, 5′-GTT CTT AAT CCC TTT CTC AAC TTC-3′ and 5′-GCA TTT GAA ACA AGC AAT GGA TAG-3′, were used for amplification of DNAs from 10 humans (two African Americans, two Caucasians, four Han Chinese, and two Ami aborigines in Taiwan), three chimpanzees, two gorillas, and one gibbon. The amplification was carried out using the following cycling parameters: initial denaturation at 94°C for 2 min, 35 cycles of denaturation at 94°C for 1 min, and primer annealing at 58°C for 1 min, followed by extension at 72°C for 1.5 min. The last cycle employed an extension time of 10 min. PCR products were purified from agarose gels with Gene Clean III elution kit (BIO 101, USA). The PCR products, which may be GPA, GPB, or GPE, were cloned into pCR2.1-TOPO vector.

In total, 29 human, nine chimpanzee, and six gorilla sequences were obtained. Sequencing reactions were carried out on both strands, often by using internal primers as well. The sequences were determined with dye terminator cycle sequencing reactions using Applied Biosystems 377A sequencer. Because there is no singleton found in the coding regions, PCR errors should not be a major concern. We also verified the cases of gene conversion by new rounds of PCR and sequencing. (New sequences described in this study are human AY297541 to AY297569, chimpanzee AY297570 to AY297578, gorilla AY297580 to AY297585, and gibbon AY297579.)

Data Analysis

To calculate the sequence divergence in introns, K_i (K_i = number of nucleotide changes per intronic site), Kimura's two-parameter model was used (Kimura 1980). For coding regions, we used Li's (1993) method to calculate K_a and K_s as stated above. K_i in general may be a better representation of the neutral rate than K_s. In addition, there are far more intronic sites than the synonymous ones in this study. Because of the closeness between human and macaque, the results are essentially the same if other methods are used (Yang 1997; Yang and Bielawski 2000). For the phylogenetic tree of figure 3, we used the neighbor-joining method (Saitou and Nei 1987) based on Kimura's two-parameter model. Figure 4 shows evidence of gene conversion, which distorts the phylogenetic relationship.

For the divergence of figure 5*a*, every sequence of human, chimpanzee, and gorilla (h/c/g) was compared with that of gibbon. For figure 5*b*, we compared sequences among h/c/g (but not within species). The total number of interspecific comparisons is therefore 489 ($29 \times 9 + 29 \times 6 + 9 \times 6$), based on 29 human, nine chimpanzee and six gorilla sequences. Intraspecific comparisons are not included here because of the presence of many nearly identical alleles. The average K_a/K_s and K_a/K_i values were calculated by averaging over K_s/K_a and K_i/K_a and taking their reciprocals. We present the harmonic mean because K_s and K_i are sometimes 0.

To identify the putative amino acid residues under positive selection and to estimate the strength of selection, Model 8 of codeml in the PAML package was used (Yang 1997; Yang et al. 2000). We chose one representative gene from each genealogical cluster for the analysis; that is, one GPA and GPB each from human, chimpanzee, and gorilla, respectively, as well as the single gene from orangutan, gibbon, and OWM. The human GPE-like sequences (fig. 3)

were not chosen because of the presence of two pseudo-exons out of the three exons (Blumenfeld et al. 1997).

For the malaria parasite, the nine genes used for the comparison with EBA-175 are STARP, CSP, AMA-1, Pfs25, RAP-1, sporozoite antigen, MSP3, Pfg27/25, and Pfs48/45. For genes with two major alleles in *P. falciparum*, such as EBA-175 and MSP3, the one that is closer to the *P. reichenowi* sequence was selected for comparison.

Results

Survey of Rapidly Evolving Genes Between Human and the Macaque Monkey

The goal is to identify fast-evolving genes between human and OWM. The analysis will then be extended to other primates and human populations. We use the standard procedure to measure the rate of coding sequence evolution (K_a and K_s; see *Materials and Methods*). In general, if K_a is significantly greater than K_s, putative positive selection is suggested.

In figure 1, K_a is plotted against K_s for the 280 genes from human and OWM. The data set likely has an overrepresentation of fast-evolving genes, perhaps due to a greater interest and effort at finding and publishing such genes by investigators. (The extrapolation from any data set to the whole genome will be plagued by possible biases in representation until the two respective genomes are nearly entirely sequenced and annotated.) In this data set, the average K_a is 2.67% and average K_s is 6.60%, their ratio being 0.405. The K_a/K_s value averaged across all 280 genes is 0.462. There is a slight and positive correlation ($r = 0.262$ and slope $= 0.242$) between K_a and K_s, as noticed by many authors in diverse taxonomic groups (Comeron and Kreitman 1998; Makalowski and Boguski 1998).

Genes above the diagonal have a K_a/K_s ratio greater than 1, and there are 26 of them. Many of these genes have an unusually small K_s, rather than a large K_a. Because smaller genes tend to experience wider fluctuations in K_s, the K_a/K_s criterion may result in the overinclusion of small genes among the fast-evolving ones. We therefore suggest a different measure, $\delta = (K_a - K_s)/\sigma_{Ks}$ where σ_{Ks} is the standard deviation of K_s. When $K_a/K_s > 1$, $\delta > 0$. In table 1, 13 genes with $\delta > 1$ are listed in the descending order of δ plus four other genes that have $\delta < 1$ but $K_a > 0.08$. The δ statistic may have an advantage over the standard K_a/K_s presentation when portraying the rate of nonsynonymous substitution among small genes.

One of the major categories in table 1 is the genes involved in immune response, including CD3G, CD59, CD3E, MSMB, interleukin 3, and interleukin 4. In addition, one of the three blood group–related genes, GYPA (= GPA), has also been shown to be involved in a pathogen interaction. Therefore, more than one-third of the genes listed in table 1 are involved in defense. Two mitochondrial respiratory complex enzymes, COX8 and NDUFC2, are fast evolving. These nuclear genes may have coevolved with mitochondria-encoded genes. The rest include two reproductive genes, PRM1 and SPINK2, one glycoprotein hormone, CGA, and one digestive enzyme, LYZ. The functions of MGC2217 and GW128 are yet to be identified. Among the top 13 entries, almost all of them have a K_s value below the average of 6.6%, with the exception of GPA (glycophorin A).

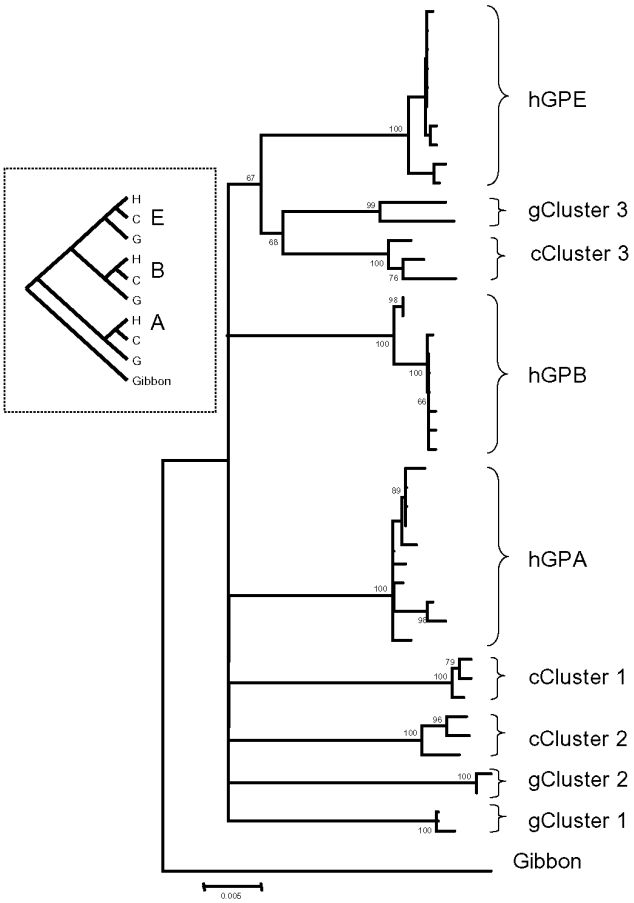

FIG. 3.—The phylogeny of the glycophorin sequences. Human sequences that conform to the canonical structure of figure 2 are labeled hGPA, hGPB, and hGPE, respectively. c and g denote chimpanzee and gorilla sequences, respectively. We designate the nonhuman sequences only as clusters 1 to 3 because the A-B-E designation is appropriate only in human (see text). Numbers at the nodes are the percentages of bootstrap support from 500 replications. Nodes with a bootstrapping value lower than 60% are collapsed into a single one. In the inset, the expected phylogeny among these genes in the absence of gene conversion is shown (Rearden et al. 1993). Note that the observed star-like phylogeny among the major clusters contrasts sharply with the expected tree of orderly bifurcation.

In our study, we chose genes for further analysis on the basis of both a high K_a/K_s ratio and a high K_a value. GPA has the highest K_a among all the genes surveyed. Another gene with a comparably high K_a value is the protamine, which has been analyzed in detail already (Rooney, Zhang, and Nei 2000; Wyckoff, Wang, and Wu 2000).

Evolution of the Glycophorins

Figure 2 shows the canonical genomic structure of GPA, GPB, and GPE in human. These sequences are defined by sites previously recognized in serological analyses and by the exon-intron splicing patterns. However, the delineation of the A, B, and E loci applies only to human. In both chimpanzee and gorilla, their GPBs (like

(a)

```
hGPA       TGGAGCCGCT GCTGGGAGGG ATGTGGGAGA GTTTGTCTTT CATAATACGC TCTATGTCCA CGCAGTCACC TCATTCTTGA CCCCTTTCTC AACTTCTCTT
hGPA-var1  ......T... .......... .......... -......... ...G...... .......... .......... .......... .......... ..........
hGPB       ......T... .......... .......... A......... ...G...... .G........ .......... .........T T......... ..........
hGPE       ......T... .......... .......... -......... ...G....A. .T........ .......... .........T T......... ..........
```

(b)

```
hGPA       GGCTTTGGTA -TAAGAGAGC TTCATGACAT AAAATGGCAA GTGGAGACAG AGACAAAAGT AGGATGTGGA CTGAGAGGGA AGGTTAGCAC AGGTGGAACA
hGPA-var-2 .....C.... G......... .......TG. .......... .......... .......... .......... .......... .......... ..........
hGPA-var-3 .....C.... G......... .......... .......... .......... .....T.... .A........ .......... .......... ..........
hGPB       .......... -......... .......... .......... .......... .......... .......... .......... .......... ..........
hGPE       .......... G......... .......TG. .......... .......... .....T.... .A........ .......... .......... ..........
```

FIG. 4.—Three examples of segmental gene conversion among human glycophorin sequences. (*a*) hGPA-var1 (variant 1 of hGPA) was converted by hGPE. The first base shown here is position 701 from the beginning of exon 2. (*b*) hGPA-var2 and hGPA-var3 were converted by hGPE. The first base shown is position 1201. These two variant sequences are from Baum, Ward, and Conway. (2002). Dot (.) and dash (-) denote identical nucleotide and gap, respectively.

human GPA) do not skip exon III. In gorilla, GPA has a common allele that skips this exon (Huang et al. 1995; Xie et al. 1997). The reason that the locus designation does not agree among species is explored below.

Gene Conversion

The 29 sequences from humans consist of 10 GPA, nine GPB, and 10 GPE alleles. The nine sequences from chimpanzee and six sequences from gorilla cannot be categorized according to the canonical structure of figure 2. Figure 3 presents the phylogeny of the glycophorin sequences from human, chimpanzee, gorilla (h/c/g for short) and the outgroup, gibbon, based on the 2-kb region shown in figure 2. Since the gene triplication occurred before the speciation among the three species (Rearden et al. 1993; Blumenfeld et al. 1997), we had expected three clusters representing the A, B, and E locus, respectively (see the inset in figure 3). Instead, nine clusters were observed. Although each cluster likely represents alleles of a locus of one species, there is no clear phylogenetic relationship among the clusters. (The appearance of an E-locus cluster is deceiving as the distances between species are too high for orthologous genes.)

A simple way to see the phylogenetic incongruence with the expectation is given in table 2. The orthologous loci between human and chimpanzee and between these two species and gorilla should be around 1.2% and 1.5%, respectively (Chen and Li 2001). Instead, no two clusters are less than 2.4% apart, suggesting the absence of true orthology among these clusters. Because the distances between all nine clusters to gibbon are close to the genomic average (Sibley and Ahlquist 1987; Li 1997), mutation rate is not the source of incongruence.

Gene conversion is a plausible explanation for the divergence patterns of figure 3 and table 2. For example, conversion of locus B by locus A in any species would destroy the former's orthology with those of the other two species. Among our human sequences, three hybrid molecules have been identified, including two MiIII variants (Huang and Blumenfeld 1991), which are the GPB allele partially converted by GPA, and one GPA allele converted by GPE in intron 3. This latter case is shown in figure 4*a*. We also noticed small segments shared by hGPA and hGPE in Baum, Ward, and Conway's (2002) sequences (figure 4*b*). In addition, more than 40 glycophorin rearrangements have been discovered (Blumenfeld and Huang 1995), and hot spots of recombination have been identified in this region (Blumenfeld and Huang 1997). Overall, sequence exchanges among loci appear quite active in these species, but the conversion has been segmental. (Otherwise, sequences of all three loci from the same species would have been more closely related phylogenetically.) Partial conversions resulted in sequences that are recombinants between loci and, hence, are of highly mixed ancestry. This mixed ancestry may account for the lack of clear-cut phylogenetic pattern in figure 3.

In general, the long-term consequence of gene conversion is homogenization among loci, but that is not the effect of our concern. The comparison we consider here is the level of genetic variation across multiple loci that undergo occasional gene conversion vis-à-vis the level of single-locus variation. The former should generally be higher than the latter. This elevated level of variation may be characterized as the "storage and retrieval" effect, which plays a central role in the "malaria evasion" hypothesis to be elaborated later.

Rate of Evolution

A result of frequent partial gene conversions is that all three loci would have evolved at a comparable rate. In figure 5*a*, the divergence between the single glycophorin of gibbon and those of all three loci of h/c/g is shown. Indeed, all h/c/g glycophorin sequences have evolved at an extraordinarily high rate. The average K_a is 0.143, three times as high as the average K_s (0.045). (In human, GPB and GPE appear to have slightly smaller K_a/K_s ratios than

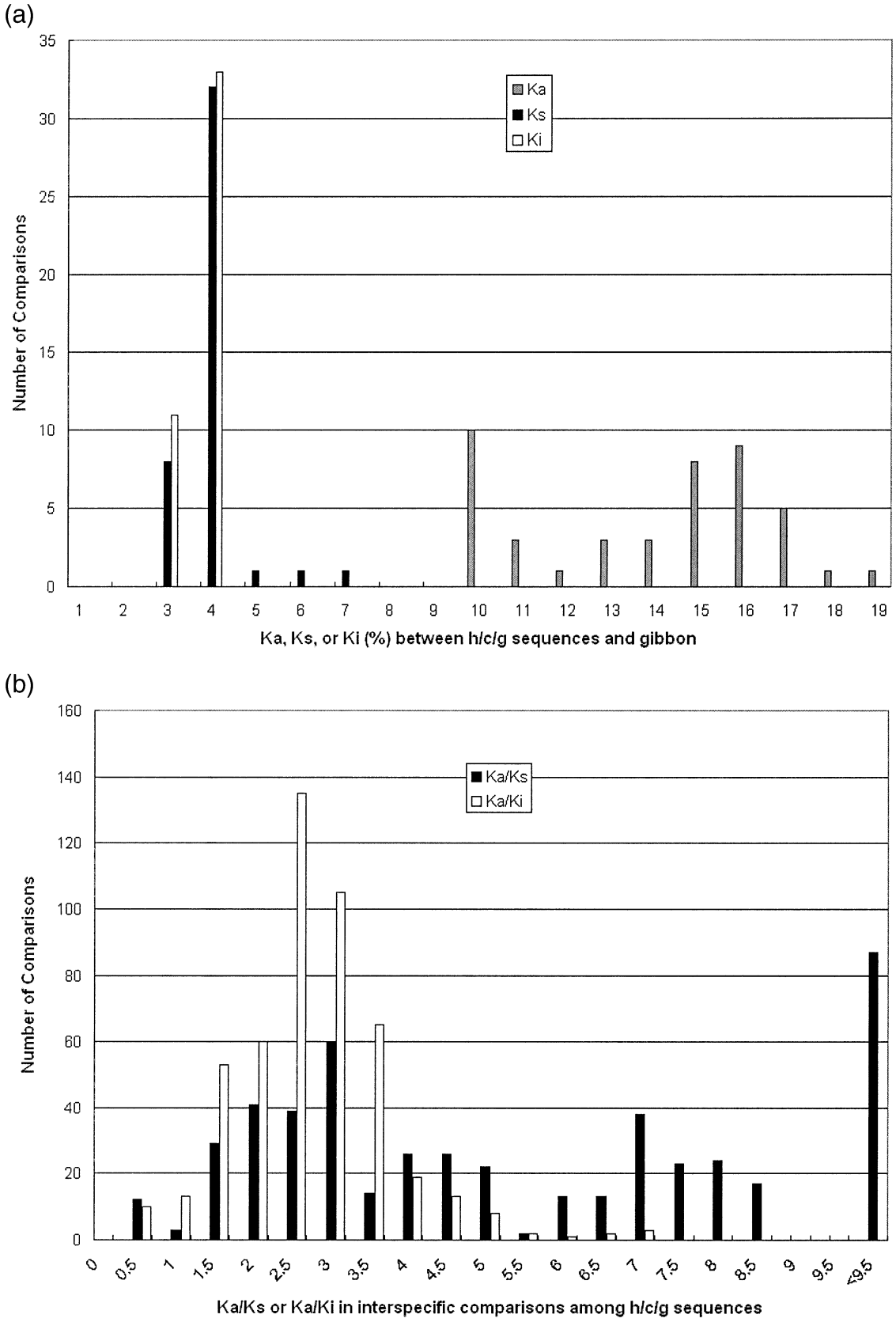

FIG. 5.—(*a*) Distributions of K_a, K_s, and K_i between the gibbon glycophorin and the 42 sequences from human, chimpanzee, and gorilla (h/c/g for short). K_a, K_s, and K_i are, respectively, the substitution numbers for nonsynonymous, synonymous, and intron sites. The results are similar when the macaque sequence is used as the outgroup. (*b*) Distributions of the K_a/K_s and K_a/K_i ratios in the interspecific comparisons between human, chimpanzee, and gorilla. Note that the number of comparisons in (*b*) is much larger than that in (*a*).

Table 1
A List of 17 Most Rapidly Evolving Genes Between Human and Old World Monkey (OWM)

Gene Name	Gene Symbol	Length	K_a	K_s	K_a/K_s	δ^a	Accession Number	
							Macaca	Human
Glycoprotein hormones, alpha polypeptide	CGA	121	0.108	0.035	3.086	3.632	ay026358	NM_000735
Protamine 1	PRM1	52	0.134	0.045	2.978	2.928	af119240	NM_002761
Lysozyme	LYZ	149	0.062	0.023	2.696	2.826	x60236	NM_000239
Glycophorin A	GYPA	151	0.152	0.066	2.359	2.527	af023469	NM_002099
Unidentified	MGC2217	86	0.079	0.028	2.821	2.277	ab072019	NM_024300
Interleukin 4	IL4	154	0.042	0.018	2.333	2.034	ab000515	NM_000589
Unidentified	GW128	64	0.048	0.016	3.000	1.939	ab072024	NM_014052
Interleukin 3	IL3	153	0.085	0.041	2.073	1.826	x51890	NM_000588
CD59 antigen	CD59	129	0.083	0.049	1.694	1.717	ab072017	NM_000611
CD55, Cromer blood group system	DAF	382	0.088	0.056	1.571	1.416	af149763	NM_000574
CD3G antigen	CD3G	183	0.095	0.059	1.610	1.295	ab073992	NM_000073
Rhesus blood group-associated glycoprotein	RHAG	429	0.060	0.041	1.463	1.226	ab015467	NM_000324
Cytochrome c oxidase subunit VIII	COX8	70	0.119	0.064	1.859	1.206	ab072015	NM_004074
CD3E antigen	CD3E	208	0.082	0.057	1.439	0.791	ab073994	NM_000733
Acrosin-trypsin inhibitor	SPINK2	85	0.128	0.103	1.243	0.639	x68331	NM_021114
NADH dehydrogenase (ubiquinone) 1, subcomplex	NDUFC2	120	0.088	0.074	1.189	0.315	ab072016	NM_004549
Microseminoprotein	MSMB	115	0.105	0.099	1.061	0.115	m92161	NM_002443

[a] $\delta = (K_a - K_s)/\sigma_{Ks}$.

GPA, an observation that may be accounted for by the presence of pseudoexons in these two genes in human.) It is striking that k_a, ranging from 0.101 to 0.196, does not overlap with k_s (0.031–0.073) in a total of 44 comparisons. The intron regions have evolved at an even slower rate than K_s, with K_i ranging from 0.038 to 0.049 (mean = 0.042).

The high rate of K_a does not depend on the choice of outgroup. Between the h/c/g sequences and that of the macaque, the average K_a/K_s ratio is 2.0, which increases to 3.56 if the outgroup is orangutan. It appears that the rate of amino acid substitution has accelerated since the apes separated from the OWMs. To address the possibility of recent acceleration in nonsynonymous substitutions, we compare the glycophorins among human, chimpanzee, and gorilla. Because each pairwise comparison represents a different genealogical depth, the K_a/K_s or K_a/K_i ratio was calculated for each of the 489 interspecific comparisons. In figure 5*b*, most of the K_a/K_s and K_a/K_i values are greater than 1, with an average of 4.0 and 2.61, respectively. These high ratios indicate that the selective pressure driving amino acid substitutions in glycophorins may have intensified since the time of African apes' common ancestor. It is significant that the increased selective pressure appears to be on all three loci.

Amino Acid Sites Under Selection

Given the large number of glycophorin coding sequences, we attempted to identify the putative amino acid residues under positive selection and estimate the strength of selection by the maximum likelihood method of (Yang et al. 2000). Using Model 8, we estimated that 78% of the residues have evolved under near neutrality with $K_a/K_s \approx 1$, and 22% have been driven by positive selection. According to this model, the average K_a/K_s ratio for these selectively driven sites is as high as 7.7. These fast-evolving sites are distinct from the glycosylation sites, which are relatively conserved among primate species (Baum, Ward, and Conway 2002). These sites are listed in Supplementary Material online.

Discussion

What drives the rapid evolution of the glycophorins? Baum, Ward, and Conway (2002) previously suggested

Table 2
Average Differences Between Sequences of the Nine Phylogenetic Clusters of Figure 3 and the Sequence of Gibbon

	hGPA	hGPB	hGPE	cCluster 1	cCluster 2	cCluster 3	gCluster 1	gCluster 2	gCluster 3
hGPA									
hGPB	0.026								
hGPE	0.028	0.031							
cCluster 1	0.028	0.035	0.036						
cCluster 2	0.027	0.033	0.033	0.030					
cCluster 3	0.027	0.028	0.029	0.033	0.032				
gCluster 1	0.024	0.037	0.034	0.034	0.024	0.032			
gCluster 2	0.031	0.033	0.031	0.037	0.034	0.026	0.033		
gCluster 3	0.025	0.035	0.033	0.031	0.027	0.033	0.024	0.031	
gibbon	0.040	0.040	0.045	0.043	0.047	0.042	0.048	0.048	0.046

NOTE.—The differences were estimated by the method of Kimura (1980).

a "decoy hypothesis" based on the rapid evolution of the GPA sequences. In this report, we show that all three glycophorin loci have comparably high K_a/K_s ratios in human, chimpanzee, and gorilla. These observations have led to an alternative suggestion, which will be referred to as the "evasion hypothesis."

The Decoy Hypothesis

In this hypothesis, GPA serves as a "decoy" to distract viruses and bacteria away from other vital organs (Baum, Ward, and Conway 2002). In general, one might not expect a decoy to evolve rapidly as it should be made easy to find. To explain the rapid evolution of GPA, the hypothesis posits that a decoy behaves like the immunoglobulins, which diversify rapidly to cope with a wide array of antigens. Glycophorins, however, have a rather different evolutionary dynamics than the immunoglobulins. Although they have been evolving rapidly, the heterozygosity in GPA in any individual is quite unremarkable, other than the common MN polymorphism. The low abundance of GPB, which underlies the Ss polymorphism, and the undetectable expression of GPE also seem incompatible with the postulate of diversity enhancement. In addition, since there are many sialic proteins specifically encoded on the erythrocyte surface (e.g., Kell, GPC, and Duffy [Schenkel-Brunner 2000]), they might be adequate decoys as well. Why then have they not been evolving rapidly like the glycophorins? Under the decoy hypothesis, pathogenic antigens must themselves be changing rapidly, so the decoy has to keep up the pace. Without identifying the candidate pathogens for which the decoy serves to distract, the hypothesis at the moment is not testable.

Alternatively, rapid evolution between loci and low diversity within locus may suggest evasion. The many incidences of interlocus conversions would also mean frequent and abrupt changes in the receptor structure. The evasion hypothesis below is very specific about the candidate pathogen and can be falsified with proper experimental setups.

The Evasion (from *P. falciparum*) Hypothesis

In human, both GPA and GPB have been shown to be the receptors of the malaria parasite, *P. falciparum* (Pasvol, Wainscoat, and Weatherall 1982; Dolan et al. 1994). The malaria ligand binding to human GPA has been identified to be the 175-kD erythrocyte-binding antigen (EBA-175). Its binding to GPA has been shown to be the primary pathway by which *P. falciparum* invades human erythrocytes (Sim et al. 1994). The proposed hypothesis is that GPA has been evolving rapidly to evade the malaria parasite. Both mutations and interlocus conversions are means of evasion; hence, the GPB and GPE sequences have been impacted by malaria as well. It should be noted that, under the evasion hypothesis, binding to EBA-175 is considered a negative pleiotropic effect of GPA. The normal function(s) of the glycophorins currently remain(s) unknown.

In parallel, EBA-175 may have been tracking the evolution of the glycophorins and, if true, should be evolving just as rapidly as the glycophorins. All these fast-evolving molecules should bear a strong signature of positive selection. In this section, we shall outline the evidence to show that the hypothesis is a viable one and deserves to be seriously tested. (The scope of the actual testing is, however, beyond the goals of the present study.)

Table 3
The Divergence of 10 Genes Between *P. falciparum* and *P. reichenowi*

	Length	K_a	K_s	K_a/K_s
Erythrocyte-binding antigen 175kD (EBA-175)[a]	4359	0.095	0.074	1.28
STARP	1896	0.058	0.046	1.26
Circumsporozoite protein (CSP)	732[b]	0.054	0.043	1.26
	1362	0.057	0.152	0.38
Apical membrane antigen-1 (AMA-1)	1578	0.049	0.065	0.71
Ookinte antigen (Pfs25)	654	0.038	0.064	0.59
FVO rhoptry–associated protein 1 (RAP-1)	2340	0.026	0.027	0.96
Sexual stage and sporozoite surface antigen	489	0.024	0.034	0.71
Merozoite surface protein-3 (MSP3)[a]	924	0.023	0.030	0.77
Pfg27/25	654	0.020	0.031	0.63
Sexual stage–specific surface antigen (Pfs48/45)	1347	0.014	0.058	0.23
Average		0.040	0.047	

[a] Only one of two major alleles was used. The other cannot be reliably aligned.
[b] This analysis excluded the central repeat region of CSP.

Interspecific Divergence in EBA-175 and Other Loci

We first examined the nucleotide substitution rate of EBA-175 vis-à-vis those of other genes between *P. falciparum* and *P. reichenowi*, the latter infecting chimpanzee (Ozwara et al. 2001). The entire EBA-175 (4,359 bp) has a K_a value of 0.095 (SE = 0.009) and a K_s value of 0.074 (SE = 0.015). The difference is significant ($P < 0.01$), as determined by the simulation method of Wyckoff, Wang, and Wu (2000). EBA-175 has apparently been under positive selection since the speciation between the two Plasmodia (and, presumably, their hosts, human and chimpanzee). We also compute the K_a and K_s values for nine other antigen-coding loci (table 3). The value ranges from 0.014 to 0.058 for K_a and 0.027 to 0.152 for K_s. Many of these antigens, including AMA-1, CSP, MSP-3, and Pfs48/45, are themselves under positive selection (Hughes and Hughes 1995; Escalante, Lal, and Ayala 1998), but none has a higher K_a value than EBA-175.

Polymorphism in EBA-175

The erythrocyte-binding domain of EBA-175 has been identified to be the 5′ cysteine-rich region, a 616–amino acid stretch called "region II" (Sim et al. 1994). DNA sequences from region II are available from 20 worldwide strains of *P. falciparum* (Liang and Sim 1997). In this sample, there are 20 nonsynonymous variants and one synonymous variant. The observed low level of synonymous polymorphism corroborates the interpretation

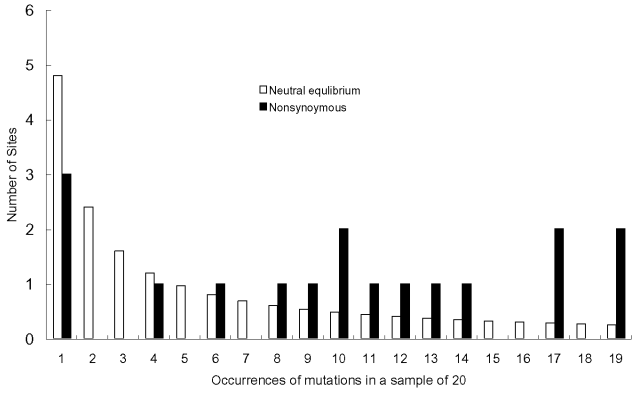

FIG. 6.—The frequency spectrum of nonsynonymous polymorphism in region II of EBA-175 of *P. falciparum*. At 17 of the 20 sites, the derived nucleotide can be inferred by reference to the *P. reichenowi* sequence; the remaining three polymorphisms are ambiguous and are not included. The frequency of occurrences of these mutations in a sample of 20 genes is given on the X-axis while the Y-axis shows the number of sites. The frequency spectrum at the neutral equilibrium is given by θ/i where i is the number of occurrences (Fu 1994). θ is the population parameter (4Nu) and is estimated by $\theta(1 + 1/2 + \cdots + 1/19) = 17$ (Watterson 1975). The lone synonymous mutation occurs only once in the sample.

of a recent loss of genetic variation in *P. falciparum* (Rich and Ayala 1998; Volkman et al. 2001).

In comparison, the level of nonsynonymous variation is too high to be attributed to demographical influences. More revealing is the population frequencies of these nonsynonymous changes. The frequency spectrum of the derived mutations is shown in figure 6. Against the neutral equilibrium (blank bar), there is an excess of high-frequency mutations by Fay and Wu's *H* statistic (Fay and Wu 2000) ($P < 0.05$), a sign of positive selection (Ewens 1979). Because *P. falciparum* is believed to have experienced a recent loss in neutral variation and should have an excess of rare mutations over the neutral equilibrium (Tajima 1989; Fu 1994), the excess in high-frequency mutations in figure 6 is even more noteworthy. Such an excess in region II can best be accounted for by either global positive selection (Ewens 1972; Fay and Wu 2000) or local selection leading to population differentiation (Fu 1994; Slatkin and Wiehe 1998). (Again, the lack of differentiation at the synonymous sites rules out the possibility of neutral population subdivision.)

Correlation Between Glycophorin Variations and the Prevalence of Malaria

The removal of GPA reduces the invasion efficiency by 50% to 95%, depending on the *P. falciparum* strain (Okoyeh, Pillai, and Chitnis 1999). The full-length extracellular domain of GPA (exons II, III, and IV) has been shown to be necessary for the binding of EBA-175 in vitro. We thus expect many naturally occurring glycophorin variants to have different binding affinity to EBA-175. An example is the En(a-) variant, which is a recombinant between GPA and GPB, and has been shown to be a poor parasite receptor in the invasion assay (Pasvol, Wainscoat, and Weatherall 1982). In addition, the specificity of *Plasmodium* invasion has been known to be very high across primate species (Escalante, Barrio, and Ayala 1995).

From structural considerations, many glycophorin variants common in regions of malaria endemics may be poor receptors for EBA-175. The He variant is a GPB epitope converted in part by GPA but with several additional mutations. The He epitope, which may make GPA GPB-like or vice versa, occurs very rarely in Caucasians and Asians but is prevalent among Africans (2% to 10%) from malaria endemic regions (Race and Sanger 1975; Mourant, Domaniewska-Sobczak, and Kopeâc 1976). In contrast, the variant St^a can reach 5% to 10% in some East Asian populations but are extremely rare among Africans and Caucasians. St^a is mostly a GPA allele that skips exon III and thus resembles GPB in the extracellular domain (Huang, Chen, and Blumenfeld 2000). A most interesting case may be the Mi-III variant, a GPB allele partially converted by GPA. The conversion restores the expression of the pseudoexon III (fig. 1), making GPB more like a variant GPA (Huang and Blumenfeld 1991). Whereas Mi-III accounts for less than 1% among Caucasians and 3% among ethnic Han Chinese, it represents 30% to 90% of GPB in several large dominant aborigines groups. These groups, especially the Ami tribe, occupied the lower elevation in Taiwan, where malaria was common in the past (Broadberry and Lin 1996).

In addition to such structural variants, many populations in regions of malaria endemics harbor unusual glycophorin variants, often in unusual frequencies. The Hunter variant can reach 22% in West Africa but is rare among Caucasians (0.5%) (Blumenfeld et al. 1997). In New Guinea, the frequency of N antigen is higher than 90%, whereas it is generally about 50% (30% to 70%) elsewhere in the world (Mourant, Domaniewska-Sobczak, and Kopeâc 1976).

Conclusions

Unlike other genetic alterations, such as sickle cell anemia or G6PD deficiency (Tishkoff et al. 2001), that confer resistance to malaria, mutations in the glycophorins would not have been debilitating even in homozygotes (Schenkel-Brunner 2000). Moreover, a reservoir of GPB and GPE variants retrievable by gene conversion or unequal exchange may produce novel GPA variants and provide human and African apes a means to evade the pursuit of pathogens. In this scenario, the advantage of gene duplication may be the ability to "store and retrieve" genetic variations. Whether (and how) glycophorins and *Plasmodium* genes interact and coevolve will have implications in public health and evolutionary theories. If this hypothesis turns out to be correct, human and ape ancestors must have been battling malaria for over 10 million years.

Pooling the evidence, we consider it plausible that the evolution of human glycophorins is at least partially driven by *P. falciparum*. It may be fruitful to systematically document the invasion efficiency of *P. falciparum* strains that carry different EBA-175 alleles. Such efficiency should be assayed against human erythrocytes carrying different glycophorin variants.

Supplementary Material

The amino acid residues that may have been driven by positive selection are listed on the journal's Web site.

Acknowledgments

We wish to thank C.-H. Huang for both the DNA samples and extensive discussions. We thank Li Jin and C. Toomajian for sharing DNA samples. We are grateful to I. Boussy, N. Osada, X. Lu, G. Morris, H. T. Yu, and S. C. Lee for the general help. H.Y.W. was supported by a predoctoral fellowship from the King Car Company of Taiwan to visit Chicago for two years. This work was also supported by NIH and NSF grants to C.-I.W. and Academia Sinica grants to J.C.S.

Literature Cited

Bamshad, M., and S. P. Wooding. 2003. Signatures of natural selection in the human genome. Nat. Rev. Genet. **4**:99–111.

Baum, J., R. H. Ward, and D. J. Conway. 2002. Natural selection on the erythrocyte surface. Mol. Biol. Evol. **19**:223–229.

Blumenfeld, O. O., and C. H. Huang. 1995. Molecular genetics of glycophorin MNS variants. Transfus. Clin. Biol. **4**:357–365.

———. 1997. Molecular genetics of the glycophorin gene family, the antigens for MNSs blood groups: multiple gene rearrangements and modulation of splice site usage result in extensive diversification. Hum. Mutat. **6**:199–209.

Blumenfeld, O. O., C. H. Huang, S. S. Xie, and A. Blancher. 1997. Molecular biology of glycophorins in humans and nonhuman primates. Pp. 113–146 *in* A. Blancher, J. Klen, and W. W. Socha, eds. Molecular biology and evolution of blood group and MHC antigens in primates. Springer, New York.

Broadberry, R. E., and M. Lin. 1996. The distribution of the MiIII (Gp.Mur) phenotype among the population of Taiwan. Transfus. Med. **6**:145–148.

Chen, F. C., and W. H. Li. 2001. Genomic divergences between humans and other hominoids and the effective population size of the common ancestor of humans and chimpanzees. Am. J. Hum. Genet. **68**:444–456.

Civetta, A., and R. S. Singh. 1998. Sex-related genes, directional sexual selection, and speciation. Mol. Biol. Evol. **15**: 901–909.

Comeron, J. M., and M. Kreitman. 1998. The correlation between synonymous and nonsynonymous substitutions in *Drosophila*: mutation, selection or relaxed constraints? Genetics **150**:767–775.

Dolan, S. A., J. L. Proctor, D. W. Alling, Y. Okubo, T. E. Wellems, and L. H. Miller. 1994. Glycophorin B as an EBA-175 independent *Plasmodium falciparum* receptor of human erythrocytes. Mol. Biochem. Parasitol. **64**:55–63.

Enard, W., M. Przeworski, S. E. Fisher, C. S. Lai, V. Wiebe, T. Kitano, A. P. Monaco, and S. Paabo. 2002. Molecular evolution of FOXP2, a gene involved in speech and language. Nature **418**:869–872.

Escalante, A. A., E. Barrio, and F. J. Ayala. 1995. Evolutionary origin of human and primate malarias: evidence from the circumsporozoite protein gene. Mol. Biol. Evol. **12**:616–626.

Escalante, A. A., A. A. Lal, and F. J. Ayala. 1998. Genetic polymorphism and natural selection in the malaria parasite *Plasmodium falciparum*. Genetics **149**:189–202.

Ewens, W. J. 1972. The sampling theory of selectively neutral alleles. Theor. Popul. Biol. **3**:87–112.

———. 1979. Mathematical population genetics. Springer-Verlag, Berlin.

Fay, J. C., and C. I. Wu. 2000. Hitchhiking under positive Darwinian selection. Genetics **155**:1405–1413.

Fay, J. C., G. J. Wyckoff, and C. I. Wu. 2001. Positive and negative selection on the human genome. Genetics **158**:1227–1234.

———. 2002. Testing the neutral theory of molecular evolution with genomic data from *Drosophila*. Nature **415**:1024–1026.

Fu, Y. X. 1994. Estimating effective population size or mutation rate using the frequencies of mutations of various classes in a sample of DNA sequences. Genetics **138**:1375–1386.

Huang, C. H., and O. O. Blumenfeld. 1991. Molecular genetics of human erythrocyte MiIII and MiVI glycophorins: use of a pseudoexon in construction of two delta-alpha-delta hybrid genes resulting in antigenic diversification. J. Biol. Chem. **266**:7248–7255.

Huang, C. H., Y. Chen, and O. O. Blumenfeld. 2000. A novel St(a) glycophorin produced via gene conversion of pseudoexon III from glycophorin E to glycophorin A gene. Hum. Mutat. **15**:533–540.

Huang, C. H., S. S. Xie, W. Socha, and O. O. Blumenfeld. 1995. Sequence diversification and exon inactivation in the glycophorin A gene family from chimpanzee to human. J. Mol. Evol. **41**:478–486.

Hughes, M. K., and A. L. Hughes. 1995. Natural selection on *Plasmodium* surface proteins. Mol. Biochem. Parasitol. **71**:99–113.

Johnson, M. E., L. Viggiano, J. A. Bailey, M. Abdul-Rauf, G. Goodwin, M. Rocchi, and E. E. Eichler. 2001. Positive selection of a gene family during the emergence of humans and African apes. Nature **413**:514–519.

Kimura, M. 1980. A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. J. Mol. Evol. **16**:111–120.

Kitano, T., and N. Saitou. 1999. Evolution of Rh blood group genes have experienced gene conversions and positive selection. J. Mol. Evol. **49**:615–626.

Li, W. H. 1993. Unbiased estimation of the rates of synonymous and nonsynonymous substitution. J. Mol. Evol. **36**:96–99.

———. 1997. Molecular evolution. Sinauer Associates, Sunderland, Mass.

Liang, H., and B. K. Sim. 1997. Conservation of structure and function of the erythrocyte-binding domain of *Plasmodium falciparum* EBA-175. Mol. Biochem. Parasitol. **84**:241–245.

Makalowski, W., and M. S. Boguski. 1998. Synonymous and nonsynonymous substitution distances are correlated in mouse and rat genes. J. Mol. Evol. **47**:119–121.

Mourant, A. E., K. Domaniewska-Sobczak, and A. C. Kopeâc. 1976. The distribution of the human blood groups and other polymorphisms. Oxford University Press, London, New York.

Okoyeh, J. N., C. R. Pillai, and C. E. Chitnis. 1999. *Plasmodium falciparum* field isolates commonly use erythrocyte invasion pathways that are independent of sialic acid residues of glycophorin A. Infect. Immun. **67**:5784–5791.

Ozwara, H., C. H. Kocken, D. J. Conway, J. M. Mwenda, and A. W. Thomas. 2001. Comparative analysis of *Plasmodium reichenowi* and *P. falciparum* erythrocyte-binding proteins reveals selection to maintain polymorphism in the erythrocyte-binding region of EBA-175. Mol. Biochem. Parasitol. **116**:81–84.

Pasvol, G., J. S. Wainscoat, and D. J. Weatherall. 1982. Erythrocytes deficiency in glycophorin resist invasion by the malarial parasite *Plasmodium falciparum*. Nature **297**:64–66.

Race, R. R., and R. Sanger. 1975. Blood groups in man. Lippincott, Philadelphia.

Rearden, A., A. Magnet, S. Kudo, and M. Fukuda. 1993. Glycophorin B and glycophorin E genes arose from the

glycophorin A ancestral gene via two duplications during primate evolution. J. Biol. Chem. **268**:2260–2267.
Rich, S. M., and F. J. Ayala. 1998. The recent origin of allelic variation in antigenic determinants of *Plasmodium falciparum*. Genetics **150**:515–517.
Rooney, A. P., J. Zhang, and M. Nei. 2000. An unusual form of purifying selection in a sperm protein. Mol. Biol. Evol. **17**:278–283.
Saitou, N., and M. Nei. 1987. The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol. Biol. Evol. **4**:406–425.
Schenkel-Brunner, H. 2000. Human blood groups. Springer-Verlag, New York.
Sibley, C. G., and J. E. Ahlquist. 1987. DNA hybridization evidence of hominoid phylogeny: results from an expanded data set. J. Mol. Evol. **26**:99–121.
Sim, B. K., C. E. Chitnis, K. Wasniowska, T. J. Hadley, and L. H. Miller. 1994. Receptor and ligand domains for invasion of erythrocytes by *Plasmodium falciparum*. Science **264**:1941–1944.
Slatkin, M., and T. Wiehe. 1998. Genetic hitch-hiking in a subdivided population. Genet. Res. **71**:155–160.
Smith, N. G., and A. Eyre-Walker. 2002. The compositional evolution of the murid genome. J. Mol. Evol. **55**:197–201.
Swanson, W. J., A. G. Clark, H. M. Waldrip-Dail, M. F. Wolfner, and C. F. Aquadro. 2001. Evolutionary EST analysis identifies rapidly evolving male reproductive proteins in *Drosophila*. Proc. Natl. Acad. Sci. USA **98**:7375–7379.
Tajima, F. 1989. The effect of change in population size on DNA polymorphism. Genetics **123**:597–601.
Ting, C. T., S. C. Tsaur, M. L. Wu, and C. I. Wu. 1998. A rapidly evolving homeobox at the site of a hybrid sterility gene. Science **282**:1501–1504.
Tishkoff, S. A., R. Varkonyi, N. Cahinhinan et al. (17 co-authors). 2001. Haplotype diversity and linkage disequilibrium at human G6PD: recent origin of alleles that confer malarial resistance. Science **293**:455–462.
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.
Watterson, G. A. 1975. On the number of segregating sites in genetical models without recombination. Theor. Popul. Biol. **7**:256–276.
Wyckoff, G. J., W. Wang, and C. I. Wu. 2000. Rapid evolution of male reproductive genes in the descent of man. Nature **403**:304–309.
Xie, S. S., C. H. Huang, M. E. Reid, A. Blancher, and O. O. Blumenfeld. 1997. The glycophorin A gene family in gorillas: structure, expression, and comparison with the human and chimpanzee homologues. Biochem. Genet. **35**: 59–76.
Yang, Z. 1997. PAML: a program package for phylogenetic analysis by maximum likelihood. Comput. Appl. Biosci. **13**:555–556.
Yang, Z., and J. P. Bielawski. 2000. Statistical methods for detecting molecular adaptation. Trends Ecol. Evol. **15**:496–503.
Yang, Z., R. Nielsen, N. Goldman, and A. M. Pedersen. 2000. Codon-substitution models for heterogeneous selection pressure at amino acid sites. Genetics **155**:431–449.

Naruya Saitou, Associate Editor

Accepted May 29, 2003