ARTICLE

Effects of Natural Selection and Gene Conversion on the Evolution of Human Glycophorins Coding for MNS Blood Polymorphisms in Malaria-Endemic African Populations

Wen-Ya Ko,[1] Kristin A. Kaercher,[1] Emanuela Giombini,[2] Paolo Marcatili,[2] Alain Froment,[3] Muntaser Ibrahim,[4] Godfrey Lema,[5] Thomas B. Nyambo,[5] Sabah A. Omar,[6] Charles Wambebe,[7] Alessia Ranciaro,[1] Jibril B. Hirbo,[1] and Sarah A. Tishkoff[1,*]

Malaria has been a very strong selection pressure in recent human evolution, particularly in Africa. Of the one million deaths per year due to malaria, more than 90% are in sub-Saharan Africa, a region with high levels of genetic variation and population substructure. However, there have been few studies of nucleotide variation at genetic loci that are relevant to malaria susceptibility across geographically and genetically diverse ethnic groups in Africa. Invasion of erythrocytes by *Plasmodium falciparum* parasites is central to the pathology of malaria. Glycophorin A (GYPA) and B (GYPB), which determine MN and Ss blood types, are two major receptors that are expressed on erythrocyte surfaces and interact with parasite ligands. We analyzed nucleotide diversity of the glycophorin gene family in 15 African populations with different levels of malaria exposure. High levels of nucleotide diversity and gene conversion were found at these genes. We observed divergent patterns of genetic variation between these duplicated genes and between different extracellular domains of *GYPA*. Specifically, we identified fixed adaptive changes at exons 3–4 of *GYPA*. By contrast, we observed an allele frequency spectrum skewed toward a significant excess of intermediate-frequency alleles at *GYPA* exon 2 in many populations; the degree of spectrum distortion is correlated with malaria exposure, possibly because of the joint effects of gene conversion and balancing selection. We also identified a haplotype causing three amino acid changes in the extracellular domain of glycophorin B. This haplotype might have evolved adaptively in five populations with high exposure to malaria.

Introduction

Longstanding struggles for survival between host and parasite are expected to leave footprints of natural selection in host and parasite genomes. Characterizing signatures of natural selection is therefore important for identifying functional variants at candidate genes involved in host-parasite interaction.[1,2] Malaria (MIM 611162) is one of the most life-threatening infectious diseases in human populations and causes approximately one million deaths per year. Ninety percent of malaria deaths occur in Africa, and the majority of them are caused by the *Plasmodium falciparum* parasite.[3] Invasion of erythrocytes by *P. falciparum* is central to the pathology of malaria. The primary step of invasion involves the binding of *P. falciparum* merozoite ligands to glycoproteins expressed on the surface of red blood cells. Glycophorin A (GYPA [MIM 111300]) and B (GYPB [MIM 111740]) are common glycoproteins on the erythrocyte surface; glycophorin A is the most highly abundant, and amino acid variation in these two proteins determines the MN and Ss blood types, respectively.[4] These glycoproteins are recognized by the erythrocyte-binding antigen 175 (EBA-175) and erythrocyte-binding ligand 1 (EBL-1), respectively; both of which are expressed by the *P. falciparum* parasite.[5–8] In addition, these two carbohydrate-rich glycoproteins supply the cell with strong negative charges thought to prevent adhesion between blood cells and vascular endothelial surfaces. The loci encoding glycophorins A and B and another another duplicated gene, glycophorin E (*GYPE* [MIM 138590]), are situated as a tandem array that spans a total of 300 kb on chromosomal region 4q31-34. These three genes are highly similar to each other at the nucleotide level (>95%), resulting in unequal recombination and gene conversion. However, they differ in the lengths of their extracellular domains because of the presence of pseudoexons in *GYPB* and *GYPE* (Figure 1A).[4]

Two prior studies have reported evidence of adaptive evolution at the three glycophorin (GYP) genes across primates and balancing selection at exon 2 of *GYPA* in a West African population.[9,10] However, the sample sizes were limited and the studies reached different conclusions about the role of glycophorin A in pathogen-receptor interaction; one of the studies suggested that *GYPA* receptors act as decoys for pathogens,[9] whereas the other study suggested that variation at *GYPA* plays a role in evading pathogen invasion.[10] Patterns of genetic variation and signatures of natural selection across ethnically diverse humans have not previously been well characterized in diverse African populations with differing exposure to malaria. Furthermore, the recent completion of the

[1]Department of Genetics and Biology, School of Medicine and School of Arts and Sciences, University of Pennsylvania, Philadelphia, PA 19104, USA; [2]Department of Biochemical Sciences "Rossi Fanelli" University of Rome "La Sapienza" P.ale Aldo Moro, 5 - 00185 Rome, Italy; [3]UMR 208, Institut de Recherche pour le Développement, Muséum National d'Histoire Naturelle, Musée de l'Homme, 75116 Paris, France; [4]Department of Molecular Biology, Institute of Endemic Diseases, University of Khartoum, 15-Khartoum, Sudan; [5]Department of Biochemistry, Muhimbili University of Health and Allied Sciences, Dar es Salaam, Tanzania; [6]Kenya Medical Research Institute, Center for Biotechnology Research and Development, 54840-00200 Nairobi, Kenya; [7]International Biomedical Research in Africa, Abuja, Nigeria
*Correspondence: tishkoff@mail.med.upenn.edu
DOI 10.1016/j.ajhg.2011.05.005.

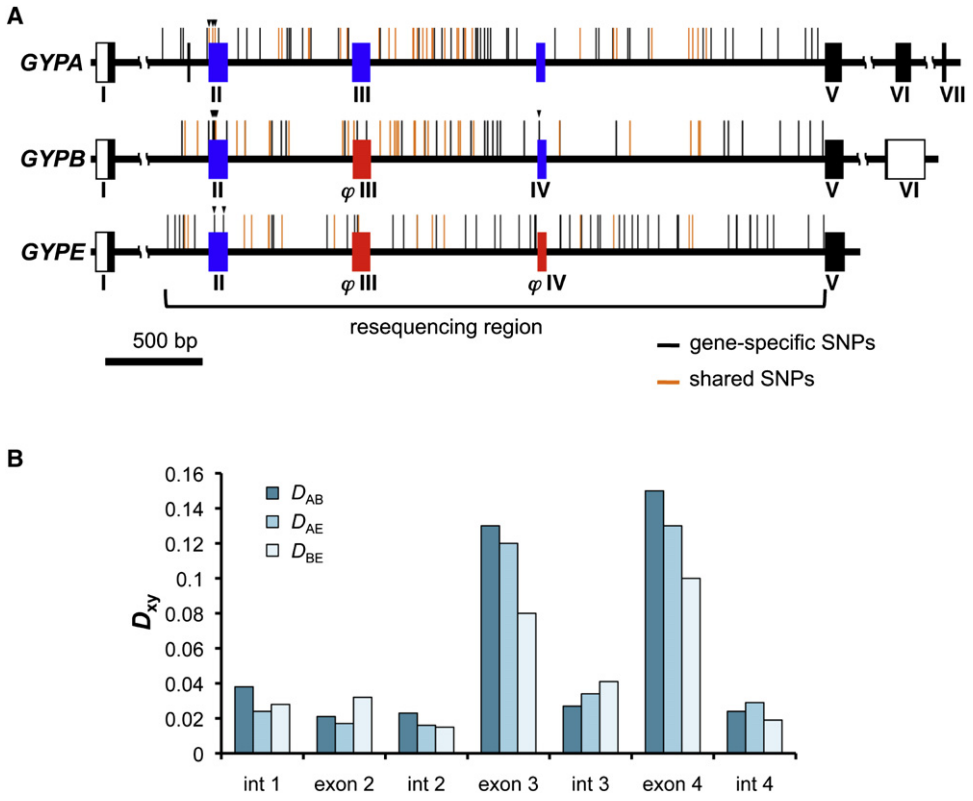

Figure 1. Spatial Distribution of SNPs and Genetic Difference between Glycophorins

(A) Gene structures of human glycophorin A (*GYPA*), B (*GYPB*), and E (*GYPE*) are marked by SNPs that are classified into "gene-specific" SNPs (black lines) or "shared" SNPs between these three paralogs (orange lines). Each nonsynonymous SNP is labeled by an inverted triangle. The exonic regions that code for the extracellular domains are colored in blue. For *GYPB* and *GYPE,* the region homologous to exon 3 or 4 of *GYPA* is colored in red if it is an unexpressed pseudoexon (φ). The exonic regions that code for the 3′ UTR in *GYPA* and *GYPE* are not shown. The three homologs are tandemly arrayed on the long arm of human chromosome 4 (cytogenetic map: 4q28.2-31.3).

(B) We used Nei's D_{xy} to estimate pairwise genetic differentiation between paralogs at different genetic regions. Nei's D_{xy} calculates the average number of "fixed" differences (per site) between any two paralogs. D_{xy} estimates of pairwise difference for the *GYPA-GYPB*, *GYPA-GYPE*, and *GYPB-GYPE* pairs are labeled as D_{AB}, D_{AE}, and D_{BE}, respectively. "int" is used as an abbreviation for intron.

chimpanzee and orangutan genome sequences enable us to better resolve the phylogeny of these duplicated genes and test models of adaptive evolution between primates and among ethnically diverse human populations.

Because of high levels of genetic structure in Africa and the possibility of local adaptation in Africans who live in distinct environments and with different pathogen exposure, we sequenced 3.7 kb of *GYPA, GYPB*, and *GYPE* in individuals from 15 African populations with differing levels of malarial endemicity.[11] We first determined orthologous loci of each GYP gene in the chimpanzee and orangutan genomes in order to infer polymorphic and fixed changes in humans. We characterized patterns of DNA polymorphism and divergence in these genes across populations and identified the underlying evolutionary mechanisms and candidate variants targeted by selection.

Material and Methods

Ethnic Groups and DNA Samples

DNA samples were from 282 unrelated individuals originating from 15 different African ethnic groups. These populations were selected on the basis of prior analyses[11] to represent genetically diverse populations and diverse geographic locations with different endemic levels of malaria (Figure 2). These populations include Yoruba from Nigeria; Bakola pygmy, Lemande, Fulani, and Mada from Cameroon; Bulala from Chad; Banuamir and Hadandawa Beja from Sudan; Borana, Boni, Sengwer, and Luo from Kenya; and Hadza, Datog, Iraqw, and Sandawe from Tanzania. Samples were collected and analyzed after approval from the institutional review boards (IRBs) of the University of Maryland and the University of Pennsylvania was obtained. Research and ethics approval and permits were obtained from the following institutions prior to sample collection: Commission for Science and Technology and the National Institute for Medical Research in Dar es Salaam, Tanzania; the Kenya Medical Research Institute in Nairobi, Kenya; the University of Khartoum in Sudan; the Nigerian Institute for Research and Pharmacological Development, Abuja, Nigeria; the Ministry of Health and National Committee of Ethics, Cameroon. Written informed consent was received from all participants. Subject identity was anonymized.

For each participant, white blood cells were isolated from whole blood by a modified salting-out procedure.[12] DNA samples were extracted in the laboratory with a Purgen DNA extraction kit (Gentra). European and Northern Chinese HapMap samples were obtained from Coriell Cell Repositories, Camden, NJ.

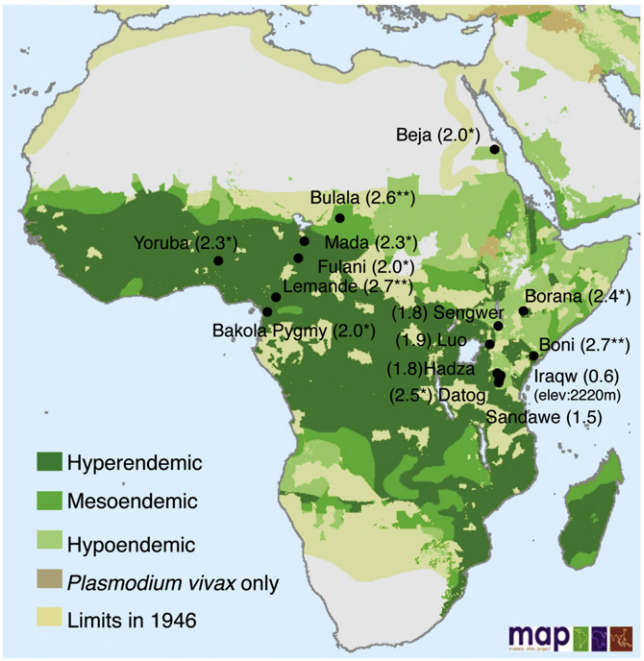

Figure 2. Geographic Distribution of Malarial Endemicity and Sampled Populations in Africa

The map of spatial limits and endemic levels of malaria in Africa was kindly provided by Drs. Robert W. Snow and Carlos Guerra from the Malaria Atlas Project published in Snow et al.[22] and was reprinted with permission from the Nature Publishing Group and Macmillan Publishers Ltd. The following endemicity classes are shown: dark green, hyperendemic and holoendemic (area in which childhood infection prevalence is $\geq$ 50%); medium green, mesoendemic (area with infection prevalence between 11%–50%); and light green, hypoendemic (area with infection prevalence $\leq$ 10%). The spatial limit for malaria transmission in 1946 is also shown. The value in parentheses indicates Tajima's *D* statistics for exon 2 of *GYPA*. Significance was calculated by coalescence simulations with no recombination for a given observed number of segregating sites (*S*) and number of chromosomes sampled in a population. *$p < 0.05$, **$p < 0.01$.

PCR and DNA Resequencing

A 3.7 kb region encompassing exons 2–4 of *GYPA* and its homologous regions from *GYPB* and *GYPE* was amplified with long-range PCR (Figure 1A). We were able to obtain full sequences for 242, 254, and 268 individuals for the *GYPA*, *GYPB*, and *GYPE* loci, respectively. Each reaction mixture contains 1× of PCR buffer, 2 mM of $MgSO_4$, 1 unit of Platinum *Taq* DNA polymerase High Fidelity (Invitrogen), 200 μM of dNTP (Promega), 0.2 μM of each primer, and 50 ng of genomic DNA (final volume, 25 μl). PCR cycles consisted of one cycle of preincubation (94°C for 1 min), 35 cycles of amplification (95°C for 30 s, 58°C for 30 s, and 68°C for 6 min 30 s), and one cycle of extension (72°C for 5 min). PCR products were prepared for sequencing with alkaline phosphatase and exonuclease I (United States Biochemicals) and sequenced by an automated ABI 3730XL sequencer (Applied Biosystems). The sequence contigs were assembled, and heterozygous sites were identified with the Sequencher 4.8 application (Gene Codes Corporation). In order to distinguish among the GYP paralogs, we designed gene-specific primers (see Table S1) that target several fixed differences among these genes, particularly at exons 3 and 4, which differ at the nucleotide level among these GYP genes by ~10%. To further confirm the specificity of each primer pair, we examined sequence trace files of each GYP sequence for an individual by aligning them across these three GYP genes. A nonspecific primer pair that amplified more than one GYP gene could be detected because all sites with mismatches between the GYP genes are expected to show multiple signal peaks on the sequence chromatograms.

Genetic Diversity and Frequency Spectra of SNPs

We aligned nucleotide sequences across all three genes by using the CLUSTAL W algorithm in the MegAlign application (DNASTAR software package) and confirmed by eye.[13] Haplotypes were inferred computationally with PHASE version 2.1.[14,15] Nei's average pairwise nucleotide difference (π) and Waterson's estimate of heterozygosity (θ_w) were calculated to estimate genetic diversity in a population.[16,17] Tajima's *D* statistic was calculated to describe allele frequency spectrum in a population by comparing the difference between π and θ_w (scaled by their covariance).[18] For a given genetic region, we generated a null distribution of Tajima's *D* by using coalescent simulations to determine the probability of an observed value under the standard neutral model for a single-copy gene, assuming no recombination and constant population size over generations given an observed number of segregating sites (*S*) and number of chromosomes sampled. We did not attempt to produce a null distribution for the test statistics by incorporating gene conversion because the model of gene conversion between these glycophorins is uncertain. Possible mechanisms that could result in a deviation from the null mode are further discussed in the Discussion section. Genetic differentiation between paralogous genes was also computed with Nei's D_{xy} estimate that calculates the average number of fixed differences between any two paralogs with a Jukes-Cantor correction for multiple hits.[16] Gaps in the sequence alignments were excluded from the analysis.

We estimated the population recombination parameter ($4Nc$, where N is effective population size and c is recombination rate per gene per generation) based on of the variance of π according to Hudson (1987).[19] Pairwise linkage disequilibrium (LD) between nucleotide variants was also computed by the squared correlation coefficient, $r_{ij}^2 = (p_{ij} - p_i p_j)^2/[p_i(1-p_i)p_j(1-p_j)]$, where p_{ij} is the frequency of one of the four possible gametes for a given pair of loci *i* and *j* and p_i and p_j are the allele frequencies of loci *i* and *j*, respectively.[20] r^2 was measured for all possible pairs of polymorphic sites except for sites that carry more than two alleles or contain gaps in the alignment. These estimates were computed with DnaSP version 5.10.[21]

Correlation between Allele Frequency Spectra and Levels of Malarial Endemicity

We calculated Spearman's rank correlation coefficient (ρ) between Tajima's *D* statistic and levels of malarial endemicity. The level of malarial endemicity for each African ethnic group was determined by mapping the latitude and longitude coordinates of the sampling site onto a global map of the spatial limit of malaria levels produced by Snow et al. (see Figure 2).[22] The estimates of Tajima's *D* at *GYPA* exon 2 as well as levels of malarial endemicity were transformed to ranks for correlation analysis for the 15 African ethnic groups and for a larger data set that also includes 29 and 30 individuals from the Northern Chinese and European HapMap populations, respectively, that currently are only rarely exposed to malaria.

Phylogenetic Analysis of *GYPA*, *GYPB*, and *GYPE*

We analyzed the phylogenetic relationship of the *GYPA*, *GYPB*, and *GYPE* glycophorin loci in humans by first identifying their

homologs in chimpanzee (*Pan troglodytes*), orangutan (*Pongo pygmaeus abelii*), and rhesus macaque (*Macaca mulatta*). Homologous regions of human *GYPA* were identified by BLAST[23] searches against the chimpanzee genome assembly (build 2.1 released in October 2006), the orangutan genome assembly (ponAbe2) released in July 2007, and the rhesus macaque genome assembly (build 1.2 released in October 2010).

We aligned all identified homologous sequences with the human *GYPA*, *GYPB*, and *GYPE* sequences (extracted from the annotated human chromosome 4 genomic contig, NW_922217) by using the CLUSTALW algorithm implemented in the MegAlign application of the DNASTAR software package. We examined the aligned sequences by eye, particularly when a gap had been introduced into the alignment. Phylogenetic trees were reconstructed with neighbor-joining and maximum likelihood algorithms and assuming the HKY85 model for nucleotide substitution and a gamma distribution for rate variation among sites as well as unweighted maximum parsimony.[24–28] We also performed a bootstrap resampling analysis with 1000 replicates to evaluate the tree topologies obtained from each algorithm.[29] The aligned sites that contain gaps were removed from the analysis. The total length of sequence alignment for phylogenetic analysis is 5618 bp. PAUP 4.0b10 was used for the tree reconstructions and bootstrap analysis.[30]

Inferring Derived Changes in the Human Lineage

To infer polymorphic and fixed derived changes on the human lineage, we reconstructed the ancestral sequences of the most recent common ancestor (MRCA) among the human sampled sequences as well as the common ancestor before the split of humans and chimpanzees by using a maximum likelihood method implemented in PAML (BASEML).[31] Because BASEML requires a single gene tree for ancestral state reconstruction, all human phased haplotype sequences were summarized onto two hypothetical sequences: for nucleotide sites that are identical among the sampled sequences, the same nucleotides were assigned to both hypothetical sequences. For polymorphic sites with two variants at a site, different variants were assigned randomly to either of the two sequences. In a coding region, codons are treated as the unit for the assignment of variants instead of nucleotides. This process summarized polymorphic sites from multiple polymorphic sequences to two hypothetical sequences regardless of their frequencies and genealogies. Polymorphic sites with more than two variants at a site were excluded from this analysis (such events are rare). Orthologous sequences from chimpanzee and orangutan were included for inferring ancestral states.

BASEML was set so that it estimated equilibrium base composition from the terminal sequences and search for the maximum likelihood estimates (MLEs) for the ratio of the rates of transitions to transversions (HKY85 substitution model) and branch lengths for given sequence data and a specified unrooted tree topology. The posterior probabilities of reconstructed ancestral states at internal nodes were determined given these MLEs. For a coding region, parameters were estimated separately for each of the three codon positions and assumed constant over a gene tree. A joint probability for a derived codon was calculated assuming independent evolution at each codon position. The posterior probability for a derived nucleotide or codon was treated as its count of changes. Derived mutations after the MRCA were treated as polymorphic changes. Mutations derived before the MRCA (and shared by all sampled human sequences) were treated as fixed changes.

Numbers of synonymous and nonsynonymous changes were determined according to Nei and Gojobori's method.[32] When a derived codon differed from its ancestral codon at more than one position, we calculated the number of synonymous and nonsynonymous changes by averaging the changes over all possible minimum-step evolutionary pathways; these pathways were weighted by their relative probabilities, which were calculated from the total counts of synonymous and nonsynonymous changes for codons that differed in only one position between derived and ancestral codons. This maximum likelihood-based approach of ancestral state reconstruction is detailed in Akashi et al.[33]

Classification of Gene-Conversion- and Mutation-Derived Changes

Gene conversion has been considered an important evolutionary mechanism for the evolution of multigene families.[34,35] To further investigate the contribution of gene conversion to nucleotide diversity in each of the three glycophorin genes, we also classified derived polymorphic and fixed changes into gene-specific or shared changes by using the method of Innan.[36] In a sequence alignment of the three glycophorins, if a derived polymorphic or fixed change in a given gene is identical to the nucleotide in either or both of its two paralogs at the same sequence position, then this derived change is classified as a shared change that was introduced by gene conversion. A derived change that does not share the same nucleotide with its two paralogs is classified as a gene-specific change. This method ignores effects of parallel mutations that could also produce shared changes between duplicated genes. However, a shared change produced from parallel mutations would require at least two independent mutation events after gene duplication, whereas only one mutation and one gene-conversion event are required for producing a shared mutation by gene conversion. Because a gene-conversion event occurs at a probability at least 100 times greater than a point mutation in the human genome,[37,38] the assumption that shared changes are due to gene conversion appears to be warranted.

Comparison of Polymorphism and Divergence for Identifying Signatures of Positive Selection

Kimura's studies on the relative contribution of weakly selected mutations to polymorphism and divergence[39] show that selection has greater effects on nucleotide divergence than on polymorphism. Hence, a comparison of the ratios of polymorphism to divergence (r_{pd}) between different classes of mutations provides a useful method for the detection of positive selection that favors mutations in one class over the other class (i.e., the McDonald Kreitman test).[40,41] We used a 2 × 2 contingency table to compare numbers of polymorphic and fixed changes for any two classes of mutations (e.g., intronic and nonsynonymous mutations). An equal ratio of polymorphism to divergence in a table indicates no fitness differences between the two classes of mutations (null hypothesis), whereas a lower r_{pd} in one class of mutations indicates greater fixation rates over the other class of mutations. We performed a *G* test for goodness of fit with William's correction to test for equal ratios of polymorphism to divergence between two given classes of mutations under the null hypothesis. When the numbers in a 2 × 2 contingency table are too small (<3), we performed a Fisher's exact test instead of a *G* test to test for homogeneity. Fisher's exact test only allows comparisons of integers; when table entries were not integers, the adjacent integers that

give a smaller r_{pd} value than the original r_{pd} were chosen for intronic changes. For nonsynonymous changes, the adjacent integers that give a larger r_{pd} value than the original r_{pd} were chosen. This is a conservative approach to detecting heterogeneity in r_{pd} because new r_{pd} values of nonsynonymous and intronic changes become closer to each other than the original values. Because the hypothesis test of homogeneity in r_{pd} was performed multiple times, we corrected for multiple testing by using the false discovery rate (FDR) procedure developed by Benjamini and Hochberg.[42]

Structural Models of GYPA and GYPB

Glycophorins A and B contain many attached sialic acids that interact directly with the *P. falciparum* ligands.[43] Because the spatial structure of sialic acids can be altered by amino acid substitutions, we attempt to illustrate relative positions of sialic acids and amino acid variants on their protein structures. Because GYP homologous sequences with known protein structures are not available in the public database, 3D structures of the proteins cannot be inferred by comparative modeling techniques. We therefore predicted the putative structures of GYPA and GYPB by assembling the short fragments of known proteins on the basis of a Monte Carlo strategy by using Rosetta.[44] We first generated 10,000 models for each protein according to the abrelax protocol adopted in CASP7 (de novo modeling followed by a full-atom relax refinement).[45] The first 1000 models with the lowest scores were clustered into different model subsets by the following criteria: (1) root mean square deviation (RMSD) threshold within members of the clusters ranging from 1.5 to 5 Å, (2) at least five models per cluster, and (3) at least 250 decoys per cluster. Finally, the lowest-scoring model in the largest cluster was selected as the candidate model. Any disordered regions evaluated by DISOPRED (protein dynamic disorder prediction program) and PSIPRED (protein secondary structure prediction based on position-specific scoring matrices) were excluded from the analysis.[46,47]

Results

Phylogenetic Analyses of Glycophorin A, B, and E Evolution in Great Apes

Loci homologous to human glycophorin A in the chimpanzee, orangutan, and rhesus macaque genome assemblies were identified with BLAST. We identified five homologous genes in the chimpanzee reference genome tandemly arrayed over approximately 490 kb on the long arm of chromosome 4 (genomic position 147.68-148.17 Mb; Figure 3). Only one homologous gene was identified in the orangutan (genomic position 149.59-149.60 Mb on chromosome 4) and in the rhesus macaque (genomic position 136.29-136.30 Mb on chromosome 5). Our analysis did not include data from the gorilla because its draft genome assembly was not yet available from the public databases by the time of manuscript submission, although multiple GYP homologs have been reported.[10,48,49] Our results are consistent with a previous study suggesting that the duplication events giving rise to *GYPB* and *GYPE* occurred after the divergence of the orangutan from the rest of the great apes about 9–13 million years ago.[48,50–52]

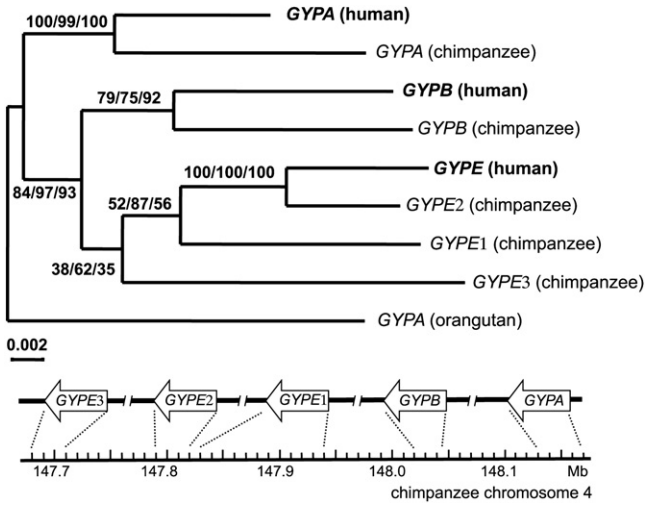

Figure 3. Phylogeny of Glycophorins in Great Apes
A maximum likelihood tree of glycophorin homologous genes in human, chimpanzee, and orangutan was reconstructed by maximizing the parameters that account for the nucleotide substitution assuming the HKY85 model, substitution-rate variation among sites assuming a gamma distribution, and branch lengths. Numbers on each internal node (from left to right) indicate scores of bootstrap resampling (1000 replicates) for maximum likelihood (ML), neighbor-joining (NJ), and maximum parsimony (MP) methods, respectively. The total length of sequence alignment is 5618 bp. The name of each chimpanzee homolog is based on its most closely related gene in humans. A number is also given if more than one chimpanzee gene is assigned to the same human ortholog.

To determine the orthologous relationships of the chimpanzee GYP homologs relative to *GYPA*, *GYPB*, and *GYPE* in humans, we performed phylogenetic analysis by using maximum likelihood (ML), neighbor-joining (NJ), and maximum parsimony (MP) algorithms. The homologous sequence identified in the macaque genome was not included in the phylogenetic analysis because it is missing large amounts of data at the region homologous to *GYPA* exons 2–3 and intron 2. A consistent tree topology was obtained by all three phylogenetic methods (Figure 3). The human *GYPA* was grouped with its orthologous locus in chimpanzee with 100%, 99%, and 100% bootstrap support for ML, NJ, and MP analyses, respectively. The gene arrangement on a chromosome was also expected to reveal the evolutionary relationship between duplicated genes.[53] Similar to the chromosomal arrangement observed in humans, the GYP homolog adjacent to *GYPA* on chimpanzee chromosome 4 clustered with human *GYPB* with 92%, 79%,and 75% bootstrap support for MP, ML, and NJ analyses, respectively. Three copies of loci homologous to *GYPE* were identified in chimpanzee, located upstream of *GYPB* and *GYPA*. We observed strong support for grouping human *GYPE* and chimpanzee *GYPE2* with 100% bootstrap support by using all three methods. The phylogenetic relationship regarding *GYPE1* and *GYPE3* is less clear but probably resulted from duplication events from an ancestral *GYPE* (Figure 3).

Table 1. Genetic Diversity of Glycophorin A, B, and E in Human Populations from Africa

	GYPA							*GYPB*							*GYPE*				
Population	***n***	***S***	S_a	θ_w	***R***	**f(M) (%)**	**f(N) (%)**	***n***	***S***	S_a	θ_w	***R***	**f(S) (%)**	**f(s) (%)**	***n***	***S***	S_a	θ_w	***R***
Cameroon																			
Fulani	28	36	3	2.5	43.2	54	46	24	33	4	2.4	0.001	25	75	34	34	0	1.4	0.6
Lemande	36	42	3	2.7	31.8	44	56	38	43	4	2.7	0.001	11	89	38	38	0	2.3	0.001
Mada	38	35	3	2.2	34.1	45	55	36	21	1	1.4	0.001	6	94	38	38	0	1.4	2.7
Bakola Pygmy	34	35	3	2.3	19.4	65	35	30	49	4	3.3	3.1	30	70	38	38	0	1.9	4.1
Chad																			
Bulala	32	33	3	2.2	22.1	41	59	30	27	1	1.8	0.001	7	93	32	32	0	1.7	4.5
Kenya																			
Boni	34	43	3	2.8	31.4	50	50	34	24	1	1.6	2.4	24	76	34	34	0	1.4	2.5
Borana	38	37	3	2.4	35.5	53	47	38	39	4	2.5	2.3	37	63	32	32	0	2.1	5.1
Luo	34	45	3	3.0	43.6	65	35	34	19	1	1.2	0.9	38	62	38	38	0	1.6	0.2
Sengwer	22	37	3	2.7	32.4	50	50	36	23	2	1.5	0.001	19	81	36	36	0	2.1	5.7
Nigeria																			
Yoruba	32	40	3	2.7	36.1	53	47	36	38	4	2.5	0.001	14	86	36	36	0	2.0	3.2
Sudan																			
Beja	34	34	3	2.2	25.2	62	38	32	17	1	1.1	0.8	25	75	34	34	0	1.7	3.1
Tanzania																			
Datog	36	49	3	3.2	49.7	56	44	34	26	1	1.7	4.2	21	79	38	38	1	2.0	4.1
Hadza	36	37	3	2.9	13.3	47	53	36	20	1	1.3	0.001	19	81	38	38	0	1.1	2.4
Iraqw	22	38	3	2.8	31.6	73	27	32	23	1	1.5	1.7	50	50	34	34	0	2.1	0.2
Sandawe	28	37	3	2.6	22.9	61	39	38	30	1	1.9	4.0	26	74	36	36	1	1.4	1.6
Total	484	83	3	3.3	35.8	54	46	508	80	5	3.2	0.2	23	77	536	72	2	2.8	3.0

The following abbreviations are used: *n*, number of chromosomes; *S*, number of segregating sites; S_a, numbers of segregating sites for nonsynonymous changes; θ_w, Waterson's estimate of heterozygosity $\times 10^{-3}$ per site; *R*, population recombination parameter per gene. Allele frequencies of MN and Ss blood groups are represented as f(M), f(N), f(S), and f(s), respectively. The length of sequence alignment of *GYPA*, *GYPB*, and *GYPE* is 3,728 bp.

Patterns of Genetic Diversity within and among Glycophorin A, B, and E Loci

A 3.7 kb region encompassing exons 2–4 and adjacent introns of *GYPA* and its homologous regions from *GYPB* and *GYPE* was sequenced in 242, 254, and 268 individuals, respectively, from 15 ethnically and geographically diverse African populations with different levels of malaria exposure. High genetic variation was observed at each of the three GYP genes across all African populations. Only 1, 2, and 1 triallelic sites were found at *GYPA*, *GYPB*, and *GYPE*, respectively. After removing these sites, we identified 83, 80, and 72 SNPs at *GYPA*, *GYPB*, and *GYPE*, respectively (Table 1). At *GYPA*, we identified three common nonsynonymous SNPs, all within exon 2. One nonsynonymous SNP occurred in a region coding for the signal peptide that is removed through the intracellular secretory pathway. The other two nonsynonymous SNPs in exon 2, which code for the MN blood type, were in complete linkage disequilibrium in all populations. The M variant, which is ancestral by comparison with chimpanzee and orangutan sequences, has serine and glycine amino acids at positions 1 and 5, respectively, of the peptide. The N variant has leucine and glutamic acid at these positions. Although prior studies of genetic variation indicate high levels of genetic structure between the African ethnic groups included in our study,[11] most populations, except for Iraqw, have similar MN allele frequencies, which range from 40% to 65% for the M variant. The estimated F_{st} across the 15 ethnic groups was 0.09 (Table 1). At *GYPB*, we identified a common nonsynonymous polymorphism that determines the Ss blood polymorphism at exon 4 in all populations and a singleton SNP that causes an amino acid change at the signal peptide in the Sengwer population. Frequencies of the Ss variants are more variable between populations compared to frequencies of the MN variants ($F_{st} = 0.46$), with frequencies ranging from 6%–50% for the S variant (methionine). In addition, we identified three nonsynonymous SNPs in exon 2 of *GYPB* at positions 1, 4, and 5 of the peptide, all

in complete linkage disequilibrium. These nonsynonymous SNPs, which change the amino acid sequence at these positions from leucine-threonine-glutamic acid to tryptophan-serine-glycine, are present at a moderate frequency in five populations: Fulani (8.3% allele frequency), Lemande (5.3%), and Bakola Pygmies from Cameroon (26.7%); Yoruba from Nigeria (5.6%); and Borana from Kenya (7.9%). At *GYPE*, we identified two nonsynonymous SNPs in exon 2 at positions 4 and 21 of the peptide, each of which occurred as a singleton.

The estimates of nucleotide diversity (θ_w) are similar among *GYPA*, *GYPB*, and *GYPE* and range from 0.0022–0.0032, 0.0011–0.0033, and 0.0011–0.0023, respectively, in the 15 populations analyzed in this study. In contrast, population recombination rates ($4Nc$) differ greatly between *GYPA* and the other two genes, and estimates range from 13.3–49.7, 0.001–4.2, and 0.001–5.7 for *GYPA*, *GYPB*, and *GYPE*, respectively (Table 1). Consequently, patterns of LD also differ considerably. Although most variants in *GYPA* are free from strong association, long blocks of strong LD were observed in *GYPB* and *GYPE* (see the Figure S1).

We determined ancestral and derived alleles for each SNP by a maximum likelihood approach by using chimpanzee and orangutan sequences as outgroups. Among the derived alleles inferred across all populations, 40.8%, 53.6%, and 28.8% in *GYPA*, *GYPB*, and *GYPE*, respectively, are identified as shared SNPs that were introduced by gene conversion from their paralogs; this suggests that gene conversion contributed substantially to the genetic diversity at these genes. These shared SNPs are scattered over the entire sequenced region in each gene (Figure 1A).

Genetic differentiation between pairs of paralogous genes was also estimated with Nei's D_{xy}. In general, low levels of genetic differentiation were observed between paralogous loci (estimates of 0.030, 0.030, and 0.027 for *GYPA*-*GYPB*, *GYPA*-*GYPE*, and *GYPB*-*GYPE* pairs, respectively). We also compared genetic differentiation between paralogous genes for introns and exons separately. All introns, as well as exon 2, showed low levels of genetic differentiation ($D_{xy} < 0.04$). By contrast, exons 3 and 4 were more than three times more divergent across paralogous loci (D_{xy} of 0.08–0.15) compared to the other regions (Figure 1B).

Allele Frequency Spectrum at the GYP Genes

We characterized the allele frequency spectra by using Tajima's *D* statistic for the exons, introns, and full 3.7 kb sequence of *GYPA*, *GYPB*, and *GYPE* for the 15 African populations. At *GYPA*, the estimates of Tajima's *D* for the full-sequenced region and for most gene regions do not deviate from the values expected under neutrality. At exon 2, however, 10 out of 15 populations showed significant positive values, indicating a frequency spectrum skewed toward an excess of intermediate-frequency alleles (Table 2). Most populations located in areas with malaria hyper- or mesoendemicity have highly significant positive Tajima's *D* values at exon 2 ranging from 1.9–2.6 (Figure 2). Most populations in the regions of hypoendemicity have lower, nonsignificant values of Tajima's *D* (with the exception of the Datog and Beja), although still positive (Tajima's *D* ranging from 0.6–1.9). With Spearman's rank correlation test, Tajima's *D* values in the African populations are significantly correlated with levels of malaria exposure (Spearman's rank coefficient, $\rho = 0.54$, $p < 0.05$). This result suggests that the degree of spectrum distortion toward an excess of intermediate-frequency alleles is correlated with malaria exposure. This correlation is specific to *GYPA* exon 2 and was not observed at other regions for any GYP genes (Figure S2). To further explore patterns of variation in populations with differing levels of malaria exposure, we sequenced *GYPA* exon 2 and calculated Tajima's *D* for 29 and 30 individuals from the HapMap Northern Chinese and European populations, respectively, who have rare malaria exposure.[22] Tajima's *D* values were 1.1 for both populations and nonsignificant, a value lower than observed in any but the Iraqw African populations. When the non-African HapMap populations are included, the correlation between Tajima's *D* and levels of malaria exposure becomes even stronger (Spearman's rank correlation coefficient, $\rho = 0.66$, $p < 0.01$).

At *GYPB* and *GYPE*, we did not observe significant deviations from neutrality in the full sequence or in any of the exons individually. However, we observed significant positive values of Tajima's *D* at *GYPB* intron 4 in five populations (Borana, Luo, Beja, Iraqw, and Sandawe) and at *GYPE* intron 3 in the Hadza. A significant negative value was also detected at *GYPE* intron 4 in the Iraqw (Table S2).

Evidence of Positive Selection on Protein Evolution in the Human Lineage

To detect evidence of positive selection on protein evolution in the human lineage since divergence from chimpanzee 4–6 million years ago,[50,52,54,55] we inferred polymorphic and fixed mutations within the human lineage by using orthologous sequences in chimpanzees and orangutans as outgroups (Figure 3). We applied the McDonald and Kreitman test for neutrality,[40] which compares ratios of polymorphism to divergence (r_{pd} = polymorphic/fixed change) between nonsynonymous and putatively neutral silent changes.[41] Adaptive nonsynonymous mutations are expected to have higher fixation rates, which would result in a smaller r_{pd} value compared to silent changes.[39–41] Use of r_{pd} has been demonstrated suitable for detecting selection in multigene families because the effects of gene conversion on r_{pd} analysis are expected to be similar across different functional classes of mutations (e.g., synonymous and nonsynonymous).[36] Because of the low number of overall synonymous changes at coding regions ($\leq$3.3 in each gene), we compared r_{pd} between nonsynonymous changes at exons 2–4 that code for the extracellular peptide and adjacent intronic changes (see Table 3). At *GYPA*, a lower r_{pd} was observed for nonsynonymous changes ($r_{pd} = 2.5/6.5 = 0.4$) than for intronic changes ($r_{pd} = 80.8/26 = 3.1$; $p = 0.014$ with

Table 2. Tajima's *D* at Different Genetics Regions of Glycophorin A in African Populations

Ethnic Group	Endemicity Class	*n*	Int1	Ex2	Int2	Ex3	Int3	Ex4	Int4	All
Cameroon										
Fulani	hyper	28	0.3	2.0*	0.1	–	0.7	–	0.9	0.9
Lemande	hyper	36	−0.9	2.6**	−0.8	–	0.2	–	−0.2	0.1
Mada	hyper	38	−1.1	2.4*	−0.2	–	0.9	–	0.7	0.8
Bakola Pygmy	hyper	34	0.6	2.0*	−1.3	–	0.0	–	0.6	0.3
Chad										
Bulala	meso	32	−1.1	2.5*	0.8	–	0.9	–	0.3	1.0
Kenya										
Boni	hyper	34	−0.9	2.6**	−0.7	–	0.9	–	−0.4	0.2
Borana	hyper	38	0.4	2.4*	0.3	–	0.8	–	0.2	0.9
Luo	hyper	34	−0.2	1.9	−0.9	–	−0.1	–	0.0	−0.1
Sengwer	hypo	22	−0.6	1.8	−0.3	–	−0.4	–	0.5	0.2
Nigeria										
Yoruba	hyper	32	-	2.4*	−0.1	–	−0.6	–	−0.3	0.0
Sudan										
Beja	hypo	34	1.5	2.0*	1.0	–	0.9	–	1.1	1.4
Tanzania										
Datog	hypo	36	−0.6	2.5*	−0.3	–	−0.3	–	0.0	0.1
Hadza	hypo	36	−1.1	1.9	−0.2	–	−0.2	–	0.2	0.2
Iraqw	hypo	22	1.1	0.6	−0.6	–	0.1	–	−0.2	−0.1
Sandawe	hypo	28	0.0	1.5	−0.8	–	0.2	–	−0.1	0.1

The following abbreviations are used: *n*, number of chromosomes; Int, intron; and Ex, exon. The following endemicity classes are used: hyperendemic (hyper), area in which childhood infection prevalence is $\geq$50%; mesoendemic (meso), area with infection prevalence between 11–50%; and hypoendemic (hypo), area with infection prevalence $\leq$10%. Tajima's *D* statistic was calculated for each genetic region in each population. Significance was calculated by performing coalescence simulations with no recombination for a given observed number of segregating sites (*S*) for a given genetic region. *$p < 0.05$, **$p < 0.01$. The length of sequence alignment of *GYPA* is 3728 bp.

a Fisher's exact test comparing 3/6 and 80/26). When we exclude exon 2, which could be a target for a different type of adaptive selection (i.e., balancing selection),[9] the value of r_{pd} remains smaller for protein changes at exons 3 and 4 (r_{pd} = 0/6.1) than for intronic changes ($p = 0.00038$ with a Fisher's exact test comparing 0/6 and 80/26), suggesting positive selection on protein evolution at exons 3 and 4.

Nucleotide evolution, however, is subject to several other genetic mechanisms that could bias our statistical tests for identifying positive selection on protein evolution. GC-biased gene conversion has been suggested to play a role in base composition evolution in mammalian genomes because nucleotide changes from A/T to G/C (denoted as "WS" where A/T and G/C are abbreviated as "W" and "S," respectively) could have higher fixation rates than the changes from the reverse direction (SW).[56–59] We therefore compared r_{pd} between protein changes at exons 3 and 4 and WS intronic changes (r_{pd} = 30.7/10.7 = 2.9) at *GYPA* and observed a smaller r_{pd} for protein changes ($p = 0.001$ for a Fisher's exact test with a comparison of 0/6 and 31/11). Second, a considerable proportion of nucleotide changes were introduced by gene conversion from paralogous loci. To further investigate the effects of gene conversion on sequence evolution and r_{pd} analysis, we tested for homogeneity in r_{pd} between shared and gene-specific changes. We observed that r_{pd} is smaller for shared changes than for gene-specific changes ($G = 4.9$, $p = 0.027$), suggesting higher fixation rates for shared changes. Therefore, to be conservative for identifying positive protein evolution, we also compared r_{pd} between protein and intronic changes introduced by gene conversion (r_{pd} = 31.6/16.6 = 1.9). r_{pd} remains smaller for protein changes than for shared intronic changes ($p = 0.0035$ for a Fisher's exact test with a comparison of 0/6 and 32/17), confirming positive selection at *GYPA* exons 3 and 4.

At the *GYPB* locus, similar to *GYPA*, most nonsynonymous fixed changes (9.1 out of 10.0) occurred in exons 3 and 4. Comparison of r_{pd} between nonsynonymous changes at exons 3 and 4 (r_{pd} = 4.0/9.1 = 0.4) and intronic changes (r_{pd} = 69.5/42.9 = 1.6) is also consistent with positive selection on protein evolution ($G = 4.5$, $p = 0.033$).

Table 3. Polymorphic and Fixed Changes in Glycophorin A, B, and E in Humans

	Syn	Intron					Nonsyn		
	Total	SW	WS	Specific	Shared	Total	Exon 2	Exons 3 and 4	Total
GYPA									
Poly	1.5	36.0	30.7	49.2	31.6	80.8	2.5	0.0	2.5
Fix	0.1	7.3	10.7	9.3	16.6	26.0	0.4	6.1	6.5
r_{pd}	–	4.9	2.9**[a]	5.3*[b]	1.9**[a]	3.1**[a]	6.3	0.0	0.4*[c]
GYPB									
Poly	2.7	30.3	25.8	30.8	38.7	69.5	5.3	4.0	9.3
Fix	0.6	18.7	18.1	18.8	24.1	42.9	0.9	9.1	10.0
r_{pd}	–	1.6	1.4	1.6	1.6	1.6*[a]	5.9	0.4	0.9
GYPE									
Poly	1.0	26.6	27.4	48.7	19.3	68.0	2.0	3.0	5.0
Fix	0.0	15.6	12.2	21.0	14.3	35.2	0.0	0.0	0.0
r_{pd}	–	1.7	2.2	2.3	1.4	1.9	–	–	–

Numbers of polymorphic (poly) and fixed (fix) nucleotide change are shown for introns and exons 2, 3, and 4. r_{pd} is the ratio of the numbers of polymorphic and fixed differences. Synonymous and nonsynonymous changes are abbreviated as Syn and Nonsyn, respectively. WS represents nucleotide changes from A/T to G/C and SW represents nucleotide changes from the reverse direction. The numbers of gene-specific (specific) and shared intronic changes (shared) are given for each GYP gene. Significance was calculated with the *G* test or Fisher exact test for a 2 × 2 contingency table between two given columns of nucleotide changes after controlling for multiple testing with Benjamini and Hochberg's FDR.[42] *$p < 0.05$, **$p < 0.01$. Significance was noted on a column of intronic changes for its comparison with the nonsynonymous changes at exons 3 and 4 unless noted otherwise. The length of sequence alignment of *GYPA*, *GYPB*, and *GYPE* is 3728 bp.

[a] Significance was given for the comparison of r_{pd} between intronic and nonsynonymous changes at exons 3 and 4.

[b] Significance was given for the comparison of r_{pd} between gene-specific and shared intronic changes at *GYPA*.

[c] Significance was given for the comparison of r_{pd} between total intronic and total nonsynonymous changes at *GYPA*.

However, r_{pd} differs significantly between intron 4 (r_{pd} = 17.0/20.3 = 0.8) and the other introns (r_{pd} = 52.6/22.6 = 2.3; $G = 6.1$, $p = 0.013$). Although r_{pd} is smaller for protein changes than for the pooled changes of introns 1–3 (G = 7.0, $p = 0.008$), the differences are not significant compared to the changes at intron 4 ($G = 0.9$, $p = 0.35$). Hence, the evidence of positive selection in *GYPB* is suggestive but not unequivocal (Table 3). Interestingly, unlike in *GYPA*, r_{pd} does not differ significantly between shared and gene-specific changes in *GYPB* and *GYPE*, suggesting that fixation rates are similar between shared and gene-specific changes in these two genes. No fixed nonsynonymous changes were inferred at *GYPE* in the human lineage indicating no sign of adaptive selection in this gene. Because hypothesis tests of homogeneity were carried out multiple times in a total of 11 comparisons, we used Benjamini and Hochberg's method[42] to adjust the p values for controlling the FDR. Differences in r_{pd} for those comparisons described above remain significant after correcting for multiple testing (see Table 3 for the significance levels after the FDR correction).

Discussion

Effects of Gene Conversion on Sequence Polymorphism and Divergence

In multigene families, gene conversion among paralogous loci often plays an important role in the introduction of genetic variation to each gene.[60] Indeed, we observed substantial effects of gene conversion contributing to nucleotide diversity at each GYP locus, consistent with Wang et al.[10] In addition, it appears that patterns of recombination rates and LD differ greatly between *GYPA* and the other two paralogs. Comparisons of the ratio of polymorphism to divergence (r_{pd}) between gene conversion and gene-specific derived changes reveal faster fixation rates for changes due to gene conversion in *GYPA*. However, this pattern is not observed for *GYPB* or *GYPE*. The striking difference in the patterns of genetic polymorphism and divergence among GYP genes probably reflects a complex mechanism of gene conversion[4] and could reflect differential purifying selection among *GYPA*, *GYPB*, and *GYPE*. The unusually high levels of genetic differentiations (D_{xy}) observed at exons 3 and 4 could result from purifying selection preventing genetic homogenization between these paralogs at these two exons compared to other regions in the loci (Figure 1B). Purifying selection would probably remove gene conversion events from a population if these events introduce variants into exon 3 or 4 of *GYPA* causing functional disruption. As a result, at *GYPA*, it is possible that only gene conversion events with short tract length could survive in a population. In contrast, long gene conversion events that introduce variants into *GYPB* or *GYPE* could continue to segregate in a population because of the relaxation of functional constraint in the recently silenced exons 3 and 4 of *GYPB* and *GYPE* (Figure 1A).[4]

Therefore, we suggest that the recent pseudogenization of exons 3 and 4 in *GYPB* and *GYPE* might have resulted in asymmetric genetic exchanges between these paralogs. Under this scenario, *GYPA* would more often experience gene-conversion events with short tract length that could inflate the estimates of recombination and reduce LD between variants, whereas gene-conversion events with long tract length could cause relatively strong LD in *GYPB* and *GYPE*. Furthermore, constant input of short sequence to *GYPA* from its paralogs could have resulted in different fixation rates for gene conversion and gene-specific mutations, whereas longer tracts of gene conversion at *GYPB* and *GYPE*, which tie adjacent variants together, could result in similar r_{pd} between gene conversion and gene-specific mutations. Future studies on modeling sequence evolution by incorporating asymmetric gene-conversion models will be useful to illustrate the striking differences in patterns of molecular evolution between the *GYPA*, *GYPB*, and *GYPE* paralogs.

Correlations between Sequence Evolution at Exon 2 of *GYPA* and Malarial Endemicity

We characterized allele frequency spectra by using Tajima's *D* and showed frequency spectra skewed toward a significant excess of intermediate-frequency alleles only at *GYPA* exon 2 in many populations (Figure 2). In duplicated genes, an excess of intermediate-frequency alleles could be attributed to several possible causes. Different haplotypes could be amplified by PCR from paralogous genes or from duplicated copies (i.e., CNV) of *GYPA* exon 2. In our experiments, we used gene-specific primer pairs that target the divergent regions of *GYPA*, *GYPB*, and *GYPE* to allow sequencing each gene unambiguously. We also performed quantitative PCR targeting exon 2 of *GYPA*, *GYPB*, and *GYPE*, independently to distinguish the possibility that the patterns of variation we are observing could be due to undetected copy-number variation. We confirmed that copy-number variation is not common in our dataset and we removed any individuals with copy-number estimates greater than one at exon 2 of each gene (Figure S3). Therefore, unintended inclusion of haplotypes from homologous or duplicated copies of *GYPA* exon 2 is unlikely to be the cause for significantly positive Tajima's *D* values. Second, demographic effects such as population subdivision or admixture could also lead to departures from neutrality in allele frequency spectrum, although the impacts should be similar across different genetic regions. However, no significant deviation from neutrality was observed except for *GYPA* exon 2; this suggests that demographic effects cannot explain the pattern observed (Table 2). Because our study contains 15 African populations that are genetically highly differentiated from each other,[11] it is also unlikely that the significant positive Tajima's *D* values at exon 2 would be observed because of random chance in measurements across diverse populations with distinct demographic histories. Finally, the different haplotypes that were introduced at GYPA exon 2 from its two paralogs as a result of gene conversion could also cause a local excess of intermediate-frequency alleles. Several derived variants at exon 2 that determine the N blood type are indeed identified as shared changes introduced by gene conversion from *GYPB*, making it difficult to distinguish from balancing selection. However, gene conversion as well as other possible causes described above would not predict a correlation between Tajima's *D* and endemicity levels of malaria as observed in our study. Lastly, the extracellular peptide encoded by *GYPA* exon 2 has been shown to be a receptor for *P. falciparum* ligand EBA-175 (discussed further below).[6,43] Together, these results suggest the possibility of joint effects of gene conversion and balancing selection for maintaining genetic variation at *GYPA* exon 2.

Role of Glycophorin A in Erythrocyte Invasion by *Plasmodium falciparum*

Gagneux and Varki[43] proposed that erythrocytes might serve as decoys during pathogen invasion. Because erythrocytes vastly outnumber other blood cells, glycan-binding viruses could enter these nonnucleated cells by interacting with the sialoglycan structures on the erythrocyte surface and thus be rendered unable to infect nucleated cells.[43] Baum and colleagues suggested that both high genetic variation and rapid rates of protein evolution at *GYPA* in humans are consistent with a decoy hypothesis whereby glycophorin A serves as a decoy receptor to attract assorted pathogens.[9] On the other hand, Wang and colleagues preferred an evasion hypothesis, which states that glycophorin A evolved adaptively to escape pathogen recognition during erythrocyte invasion.[10]

Our study confirms Baum et al.'s report[9] of an excess of intermediate-frequency alleles at *GYPA* exon 2 in the Yoruba population. However, high genetic variation and rapid rates of protein evolution are found at different parts of the extracellular peptide of GYPA. We observed a frequency spectrum skewed toward an excess of intermediate-frequency alleles at exon 2 in most African populations but with little protein divergence in the human lineage (<1 nonsynonymous fixed change; see Table 3). In contrast, we detected evidence for positive selection resulting in an excess of fixed amino acid substitutions specific to the human lineage and no polymorphism at exons 3 and 4. This pattern is not consistent with the decoy hypothesis that would normally predict, at a given locus, both high genetic diversity and fast evolutionary rates, driven by diversifying selection such that the efficiency of recognizing foreign ligands is improved (e.g., antigen-recognition sites of MHC molecules[61,62]). The decoy hypothesis also cannot easily explain the role of glycophorins on erythrocyte invasion by *Plasmodium* parasites because these malaria-causing protozoans do not rely on nuclei of host cells for their reproduction.[43] Therefore, the role of glycophorin A during erythrocyte invasion by *P. falciparum* might be better explained by the antagonistic

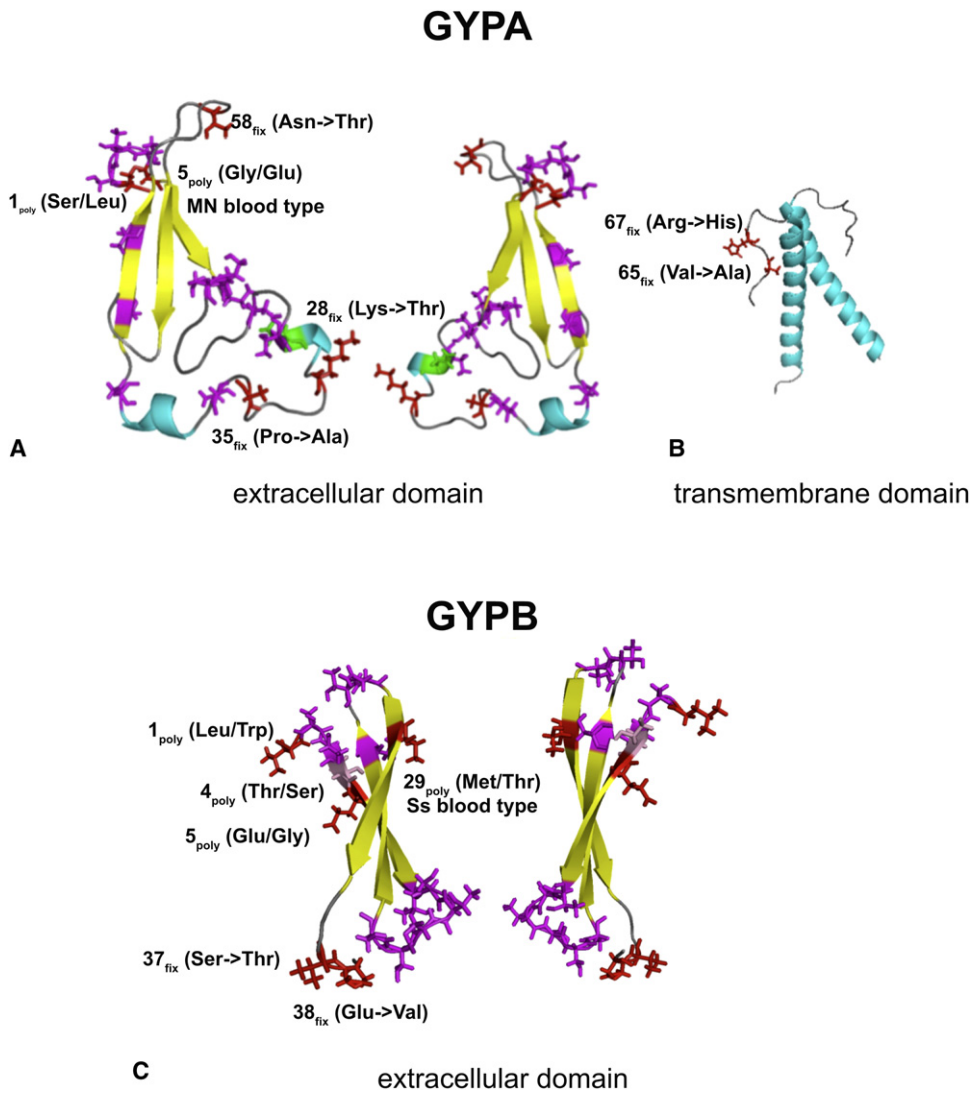

Figure 4. Structural Models of Glycophorins A and B

(A) Extracellular domain of GYPA. O- and N-glycosylated amino acid residues are labeled in magenta and green, respectively. Amino acid residues where polymorphic (poly) or fixed (fix) changes occurred are colored in red. Amino acid abbreviations at these sites are given in parentheses and the ancestral residue is shown at left and the derived one at right for a fixed change. The two images are rotated by 180 degrees with respect to each other.

(B) 3D structure of the transmembrane domain of GYPA (PDB code: 1AFO). The two fixed mutations occur at the extracellular peptide residues (in red) adjacent to the helical region.

(C) Extracellular domain of GYPB. Amino acid residues that are both glycosylated (O-linked) and variable are colored in pink. The MN and Ss blood groups are determined by amino acid variants at GYPA and GYPB, respectively. The M and N antigens are determined by two genetically linked amino acid variants of GYPA where M blood type is determined by Ser and Gly at position 1 and 5, respectively, and N blood type is determined by Leu and Glu (introduced by gene conversion from *GYPB*) at these two sites. The S and s are determined by Met and Thr, respectively, at position 29 of the polypeptide chain of GYPB.

relationship that occurs during coevolution between receptors and pathogens (i.e., the evasion hypothesis).

The divergent patterns of sequence evolution on different parts of the extracellular peptide could be related to the uneven distribution of O-sialoglycans that are mostly located at the first 26 amino acid residues (encoded by exon 2) at the NH_2 terminus of the extracellular peptide rather than at the other residues (encoded by exons 3 and 4). Whereas mutations that occur at the O-sialoglycan-rich regions are likely to alter the spatial arrangement of sialoglycan and modify binding affinity to the *P. falciparum* ligands,[43,63] mutations that occur at the O-sialoglycan-poor regions might not necessarily be selected for changing ligand-binding affinity but rather for maintaining stability of protein structure (Figure 4A and 4B).[64,65]

Tests for Selection at *GYPB* and *GYPE*

Using the McDonald and Kreitman test, we detected a signature of adaptive evolution at exons 3 and 4 of *GYPB* when we excluded intron 4 from the analysis.

Additionally, we identified a haplotype that contains three nonsynonymous mutations at *GYPB* exon 2 in five populations with high malaria exposure (see Figure 4C). Although the Tajima's *D* statistic was not significant, we performed Kelly's Z_{nS} test, a test that is more sensitive for detecting recent selection events, to detect signatures of natural selection that causes elevation of LD.[66] Significant departures from neutrality were detected for all five populations at *GYPB* exon 2 (Table S3). Although gene conversion alone could also elevate LD and lead to a significant Z_{nS} result, the same haplotype was not found in the paralogous genes *GYPA* and *GYPE*, implying that gene conversion cannot fully explain the complete LD between the three nonsynonymous mutations and that this haplotype might have evolved adaptively. Interestingly, two of the three amino acid changes in the *GYPB* peptide appear to occur at the same positions of the *GYPA* peptide at which the M/N variants are located. Furthermore, the N variant at *GYPA* and one of the three nonsynonymous mutations of this identified haplotype at *GYPB* were classified as gene-conversion-derived changes, demonstrating the effects of gene conversion on facilitating the creation of novel haplotypes upon which natural selection could act.

No evidence of adaptive protein evolution was observed at *GYPE*. Gene expression of *GYPE* on the erythrocyte surface is also undetectable,[67] suggesting that *GYPE* might not be involved in antigen-receptor interaction during erythrocyte invasion of malaria parasites. Future studies of the binding affinity of the *GYPB* variant as well as the variants that determine M/N blood types to the corresponding parasite ligands in different *P. falciparum* strains will be important for determining their possible roles on malaria susceptibility.

Supplemental Data

Supplemental Data include three figures and four tables and can be found with this article online at http://www.cell.com/AJHG/.

Acknowledgments

We are grateful to Andrew Clark, Charla Lambert, Joseph Lachance, and Laura Scheinfeldt for their insightful comments on the manuscript. We thank Bill Beggs and Karuna Panchapakesan for their technical assistance. We also thank Karen Carleton at the University of Maryland for her valuable suggestions on designing the real-time PCR experiments. We are grateful to Cheng-Han Huang at New York Blood Center for generously providing MM and MN blood-type DNA samples. We thank Ajit Varki for his valuable comments on the biology of sialic acids. This research is funded by Human Frontiers in Science grant 050913-8931, National Science Foundation grants BCS 0196183 and BCS-0827436, and National Institutes of Health grants R01GM076637 and DP1-OD-006445-01 to S.A.T.

Received: February 25, 2011
Revised: April 22, 2011
Accepted: May 5, 2011
Published online: June 9, 2011

Web Resources

The URLs for data presented herein are as follows:

Malaria Atlas Project, www.map.ox.ac.uk
National Center for Biotechnology Information (NCBI) Databases, http://blast.ncbi.nlm.nih.gov/
Online Mendelian Inheritance in Man (OMIM), http://www.omim.org

Accession Numbers

The GenBank accession numbers for the *GYP* sequences reported in this paper are HQ401724–HQ402487.

References

1. Woolhouse, M.E., Webster, J.P., Domingo, E., Charlesworth, B., and Levin, B.R. (2002). Biological and biomedical implications of the co-evolution of pathogens and their hosts. Nat. Genet. *32*, 569–577.
2. Escalante, A.A., Cornejo, O.E., Rojas, A., Udhayakumar, V., and Lal, A.A. (2004). Assessing the effect of natural selection in malaria parasites. Trends Parasitol. *20*, 388–395.
3. World Health Organization. (2008). World malaria report. 190.
4. Blumenfeld, O.O., and Huang, C.H. (1995). 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.
5. Pasvol, G., Wainscoat, J.S., and Weatherall, D.J. (1982). Erythrocytes deficiency in glycophorin resist invasion by the malarial parasite *Plasmodium falciparum*. Nature *297*, 64–66.
6. Sim, B.K., Chitnis, C.E., Wasniowska, K., Hadley, T.J., and Miller, L.H. (1994). Receptor and ligand domains for invasion of erythrocytes by *Plasmodium falciparum*. Science *264*, 1941–1944.
7. Denomme, G.A. (2004). The structure and function of the molecules that carry human red blood cell and platelet antigens. Transfus. Med. Rev. *18*, 203–231.
8. Mayer, D.C., Cofie, J., Jiang, L., Hartl, D.L., Tracy, E., Kabat, J., Mendoza, L.H., and Miller, L.H. (2009). Glycophorin B is the erythrocyte receptor of *Plasmodium falciparum* erythrocyte-binding ligand, EBL-1. Proc. Natl. Acad. Sci. USA *106*, 5348–5352.
9. Baum, J., Ward, R.H., and Conway, D.J. (2002). Natural selection on the erythrocyte surface. Mol. Biol. Evol. *19*, 223–229.
10. Wang, H.Y., Tang, H., Shen, C.K.J., and Wu, C.I. (2003). Rapidly evolving genes in human. I. The glycophorins and their possible role in evading malaria parasites. Mol. Biol. Evol. *20*, 1795–1804.
11. Tishkoff, S.A., Reed, F.A., Friedlaender, F.R., Ehret, C., Ranciaro, A., Froment, A., Hirbo, J.B., Awomoyi, A.A., Bodo, J.M., Doumbo, O., et al. (2009). The genetic structure and history of Africans and African Americans. Science *324*, 1035–1044.
12. Miller, S.A., Dykes, D.D., and Polesky, H.F. (1988). A simple salting out procedure for extracting DNA from human nucleated cells. Nucleic Acids Res. *16*, 1215.
13. Higgins, D.G., and Sharp, P.M. (1988). CLUSTAL: A package for performing multiple sequence alignment on a microcomputer. Gene *73*, 237–244.

14. Stephens, M., Smith, N.J., and Donnelly, P. (2001). A new statistical method for haplotype reconstruction from population data. Am. J. Hum. Genet. *68*, 978–989.
15. Stephens, M., and Donnelly, P. (2003). A comparison of bayesian methods for haplotype reconstruction from population genotype data. Am. J. Hum. Genet. *73*, 1162–1169.
16. Nei, M. (1987). Molecular evolutionary genetics (New York: Columbia University Press).
17. Watterson, G.A. (1975). On the number of segregating sites in genetical models without recombination. Theor. Popul. Biol. *7*, 256–276.
18. Tajima, F. (1989). Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. Genetics *123*, 585–595.
19. Hudson, R.R. (1987). Estimating the recombination parameter of a finite population model without selection. Genet. Res. *50*, 245–250.
20. Hill, W.G., and Robertson, A. (1968). Linkage disequilibrium in finite populations. Theor. Appl. Genet. *38*, 226–231.
21. Librado, P., and Rozas, J. (2009). DnaSP v5: A software for comprehensive analysis of DNA polymorphism data. Bioinformatics *25*, 1451–1452.
22. Snow, R.W., Guerra, C.A., Noor, A.M., Myint, H.Y., and Hay, S.I. (2005). The global distribution of clinical episodes of Plasmodium falciparum malaria. Nature *434*, 214–217.
23. Altschul, S.F., Gish, W., Miller, W., Myers, E.W., and Lipman, D.J. (1990). Basic local alignment search tool. J. Mol. Biol. *215*, 403–410.
24. Fitch, W.M. (1971). Toward defining the course of evolution: Minimum change for a specific tree topology. Syst. Zool. *20*, 406–416.
25. Felsenstein, J. (1981). Evolutionary trees from DNA sequences: A maximum likelihood approach. J. Mol. Evol. *17*, 368–376.
26. Hasegawa, M., Kishino, H., and Yano, T. (1985). Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. J. Mol. Evol. *22*, 160–174.
27. Saitou, N., and Nei, M. (1987). The neighbor-joining method: A new method for reconstructing phylogenetic trees. Mol. Biol. Evol. *4*, 406–425.
28. Yang, Z. (1994). Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: Approximate methods. J. Mol. Evol. *39*, 306–314.
29. Felsenstein, J. (1985). Confidence limits on phylogenies: An approach using the bootstrap. Evolution *39*, 783–791.
30. Swofford, D.L. (2003). PAUP*. Phylogenetic Analysis Using Parsimony (*and Other Methods). Version 4 (Sunderland, Massachusetts: Sinauer Associates).
31. Yang, Z. (1997). PAML: A program package for phylogenetic analysis by maximum likelihood. Comput. Appl. Biosci. *13*, 555–556.
32. Nei, M., and Gojobori, T. (1986). Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. Mol. Biol. Evol. *3*, 418–426.
33. Akashi, H., Ko, W.Y., Piao, S., John, A., Goel, P., Lin, C.F., and Vitins, A.P. (2006). Molecular evolution in the *Drosophila melanogaster* species subgroup: Frequent parameter fluctuations on the timescale of molecular divergence. Genetics *172*, 1711–1726.
34. Ohta, T. (2000). Mechanisms of molecular evolution. Philos. Trans. R. Soc. Lond. B Biol. Sci. *355*, 1623–1626.
35. Innan, H., and Kondrashov, F. (2010). The evolution of gene duplications: Classifying and distinguishing between models. Nat. Rev. Genet. *11*, 97–108.
36. Innan, H. (2003). The coalescent and infinite-site model of a small multigene family. Genetics *163*, 803–810.
37. Frisse, L., Hudson, R.R., Bartoszewicz, A., Wall, J.D., Donfack, J., and Di Rienzo, A. (2001). Gene conversion and different population histories may explain the contrast between polymorphism and linkage disequilibrium levels. Am. J. Hum. Genet. *69*, 831–843.
38. Padhukasahasram, B., Marjoram, P., and Nordborg, M. (2004). Estimating the rate of gene conversion on human chromosome 21. Am. J. Hum. Genet. *75*, 386–397.
39. Kimura, M. (1983). The Neutral Theory of Molecular Evolution (Cambridge, UK: Cambridge University Press).
40. McDonald, J.H., and Kreitman, M. (1991). Adaptive protein evolution at the Adh locus in Drosophila. Nature *351*, 652–654.
41. Akashi, H. (1995). Inferring weak selection from patterns of polymorphism and divergence at "silent" sites in Drosophila DNA. Genetics *139*, 1067–1076.
42. Benjamini, Y., and Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. J R Stat Soc Series B Stat Methodol *57*, 289–300.
43. Gagneux, P., and Varki, A. (1999). Evolutionary considerations in relating oligosaccharide diversity to biological function. Glycobiology *9*, 747–755.
44. Das, R., and Baker, D. (2008). Macromolecular modeling with rosetta. Annu. Rev. Biochem. *77*, 363–382.
45. Das, R., Qian, B., Raman, S., Vernon, R., Thompson, J., Bradley, P., Khare, S., Tyka, M.D., Bhat, D., Chivian, D., et al. (2007). Structure prediction for CASP7 targets using extensive all-atom refinement with Rosetta@home. Proteins *69* (*Suppl 8*), 118–128.
46. McGuffin, L.J., Bryson, K., and Jones, D.T. (2000). The PSIPRED protein structure prediction server. Bioinformatics *16*, 404–405.
47. Ward, J.J., Sodhi, J.S., McGuffin, L.J., Buxton, B.F., and Jones, D.T. (2004). Prediction and functional analysis of native disorder in proteins from the three kingdoms of life. J. Mol. Biol. *337*, 635–645.
48. Rearden, A., Magnet, A., Kudo, S., and Fukuda, M. (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.
49. Xie, S.S., Huang, C.H., Reid, M.E., Blancher, A., and Blumenfeld, O.O. (1997). The glycophorin A gene family in gorillas: Structure, expression, and comparison with the human and chimpanzee homologues. Biochem. Genet. *35*, 59–76.
50. Glazko, G.V., and Nei, M. (2003). Estimation of divergence times for major lineages of primate species. Mol. Biol. Evol. *20*, 424–434.
51. Hobolth, A., Dutheil, J.Y., Hawks, J., Schierup, M.H., and Mailund, T. (2011). Incomplete lineage sorting patterns among human, chimpanzee, and orangutan suggest recent orangutan speciation and widespread selection. Genome Res. *21*, 349–356.
52. Stauffer, R.L., Walker, A., Ryder, O.A., Lyons-Weiler, M., and Hedges, S.B. (2001). Human and ape molecular clocks and constraints on paleontological hypotheses. J. Hered. *92*, 469–474.

53. Iida, K., Cox-Foster, D.L., Yang, X., Ko, W.Y., and Cavener, D.R. (2007). Expansion and evolution of insect GMC oxidoreductases. BMC Evol. Biol. *7*, 75.
54. Chen, F.C., and Li, W.H. (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.
55. Hobolth, A., Christensen, O.F., Mailund, T., and Schierup, M.H. (2007). Genomic relationships and speciation times of human, chimpanzee, and gorilla inferred from a coalescent hidden Markov model. PLoS Genet. *3*, e7.
56. Eyre-Walker, A. (1993). Recombination and mammalian genome evolution. Proc. Biol. Sci. *252*, 237–243.
57. Duret, L., and Arndt, P.F. (2008). The impact of recombination on nucleotide substitutions in the human genome. PLoS Genet. *4*, e1000071.
58. Duret, L., and Galtier, N. (2009). Biased gene conversion and the evolution of mammalian genomic landscapes. Annu. Rev. Genomics Hum. Genet. *10*, 285–311.
59. Ratnakumar, A., Mousset, S., Glémin, S., Berglund, J., Galtier, N., Duret, L., and Webster, M.T. (2010). Detecting positive selection within genomes: The problem of biased gene conversion. Philos. Trans. R. Soc. Lond. B Biol. Sci. *365*, 2571–2580.
60. Ohta, T. (2010). Gene Conversion and Evolution of Gene Families: An Overview. Genes *1*, 349–356.
61. Hughes, A.L., and Nei, M. (1988). Pattern of nucleotide substitution at major histocompatibility complex class I loci reveals overdominant selection. Nature *335*, 167–170.
62. Hughes, A.L., and Nei, M. (1989). Nucleotide substitution at major histocompatibility complex class II loci: Evidence for overdominant selection. Proc. Natl. Acad. Sci. USA *86*, 958–962.
63. Jentoft, N. (1990). Why are proteins O-glycosylated? Trends Biochem. Sci. *15*, 291–294.
64. Ohta, T. (1992). The nearly neutral theory of molecular evolution. Annu. Rev. Ecol. Syst. *23*, 263–286.
65. Hartl, D.L., and Taubes, C.H. (1996). Compensatory nearly neutral mutations: Selection without adaptation. J. Theor. Biol. *182*, 303–309.
66. Kelly, J.K. (1997). A test of neutrality based on interlocus associations. Genetics *146*, 1197–1206.
67. Anstee, D.J. (1990). The nature and abundance of human red cell surface glycoproteins. J. Immunogenet. *17*, 219–225.