Published in final edited form as:

Hum Genet. 2018 February ; 137(2): 151–160. doi:10.1007/s00439-018-1866-3.

Complex signatures of natural selection at *GYPA*

Abigail W. Bigham¹, Kevin Magnaye², Diane M. Dunn³, Robert B. Weiss³, and Michael Bamshad⁴

¹Department of Anthropology, The University of Michigan, Ann Arbor, MI, USA

²Department of Human Genetics, The University of Chicago, Chicago, IL, USA

³Department of Human Genetics, The University of Utah, Salt Lake City, UT, USA

⁴Departments of Pediatrics and Genome Sciences, The University of Washington, Seattle, WA, USA

Abstract

The human MN blood group antigens are isoforms of glycophorin A (GPA) encoded by the gene, *GYPA*, and are the most abundant erythrocyte sialoglycoproteins. The distribution of MN antigens has been widely studied in human populations yet the evolutionary and/or demographic factors affecting population variation remain elusive. While the primary function of GPA is yet to be discovered, it serves as the major binding site for the 175-kD erythrocyte-binding antigen (EB-175) of the malarial parasite, *Plasmodium falciparum*, a major selective pressure in recent human history. More specifically, exon two of *GYPA* encodes the receptor binding ligand to which *P. falciparum* binds. Accordingly, there has been keen interest in understanding what impact, if any, natural selection has had on the distribution of variation in *GYPA* and exon two in particular. To this end, we resequenced *GYPA* in individuals sampled from both *P. falciparum* endemic (sub-Saharan Africa and South India) and non-endemic (Europe and East Asia) regions of the world. Observed patterns of variation suggest that *GYPA* has been subject to balancing selection in populations living in malaria endemic areas and in Europeans, but no such evidence was found in samples from East Asia, Oceania, and the Americas. These results are consistent with malaria acting as a selective pressure on *GYPA*, but also suggest that another selective force has resulted in a similar pattern of variation in Europeans. Accordingly, *GYPA* has perhaps a more complex evolutionary history wherein on a global scale, spatially varying selective pressures have governed its natural history.

Keywords

GYPA; Malaria; Balancing Selection; Human Evolution

Corresponding Author: Abigail Bigham, Department of Anthropology, University of Michigan, 222C West Hall, 1085 S. University, Ann Arbor, MI 48109-1107, USA, awbigham@umich.edu, Phone: (814) 769-9611, Fax: (734) 763-6077.

Conflict of Interest: The authors declare that they have no conflict of interest.

Electronic Supplementary Material

Supplementary Data File S1 and Supplementary Tables S1, S2, S3, S4, S5, S6 and S7 as well as DNA sequencing text files are available at Human Genetics online.

Introduction

Infectious diseases have been a major selective force during the evolution and differentiation of human populations (Bamshad and Wooding 2003; Barreiro and Quintana-Murci 2010). Many of the genes exhibiting the strongest signatures of positive selection (e.g. elevated Ka/Ks ratios, elevated dN/dS ratios, high F_{ST}) in the human lineage encode proteins involved in immunity (Nielsen et al. 2007), and changes in patterns of polymorphism in response to specific pathogens have been documented for multiple loci (Saunders et al. 2002; Verrelli et al. 2002; Wood et al. 2005). Yet, much remains to be learned about the impact of pathogens on patterns of human genetic variation.

A common means by which pathogens infect humans is to usurp cell surface glycoproteins that serve as receptors for endogenous ligands (Karlsson 1995). One canonical family of cell-surface receptors that appear to have been exploited in such a way is the glycophorins, encoded by three closely linked paralogous genes (*GYP A*, *GYP B*, and *GYP E*) that have arisen through gene duplication (Onda et al. 1994; Rearden et al. 1993). Located in tandem on the long arm of chromosome 4q28–31, *GYP A*, *GYP B*, and *GYP E* are some of the fastest evolving genes in the human lineage (Rearden et al. 1993; Wang et al. 2003). *GYP A* and *GYP B* encode GPA and GPB, respectively. Variants of GPA represent M and N antigens of the MNS blood group system and GPB the S antigen.

A high non-synonymous substitution rate has been observed for *GYP A* in humans and its orthologs in chimp, gorilla, orangutan, gibbon, and macaque (Baum et al. 2002; Wang et al. 2003). The evolutionary forces that have affected *GYP A* have been of particular interest because GPA is the major receptor to which *Plasmodium falciparum*, the most virulent species of malaria infecting humans, binds to gain cell entry. Specifically, the 175-kD erythrocyte binding antigen (EBA-175) of *P. falciparum* recognizes and docks to a receptor binding ligand on GPA (Camus and Hadley 1985; Chitnis and Miller 1994; Sim et al. 1994; Tolia et al. 2005). However, while the GPA-dependent pathway is the major means by which *P. falciparum* invades human cells, haematologically normal individuals lacking both GPA and GPB on their erythrocyte surface can be infected with *P. falciparum* via a GPA-independent pathway (Okoyeh et al. 1999; Pasvol et al. 1982).

Two hypotheses have been proposed to explain the rapid evolution of *GYP A*. Baum et al. proposed that GPA acts as a “decoy” receptor, whereby pathogens bind to GPA on an erythrocyte instead of binding to their target cells (Baum et al. 2002). This hypothesis was based on two observations. First, GPA binds multiple infectious agents including viruses and bacteria (Baseman et al. 1982; Brooks et al. 1989; Nishimura et al. 1988; Paul 1987; Saada et al. 1991; Tavakkol and Burness 1990; Wybenga et al. 1996). Second, the erythrocyte is an optimal decoy because it lacks a nucleus. Therefore, invading viruses do not have access to the cellular machinery necessary for replication (Gagneux and Varki 1999). In contrast, Wang et al. (2003) posited that *GYP A* has co-evolved with EBA-175 based on relatively weak evidence that both *GYP A* and the gene that encodes EBA-175 appear to have been subject to recent positive selection. More recently, Ko et al. (Ko et al. 2011) found a significant correlation between malaria exposure and a skew in the distribution of intermediate frequency alleles observed among populations from sub-Saharan Africa. They

suggested that this observation was best explained by a combination of balancing selection and gene conversion acting on *GYPA*.

Results

We investigated the global pattern of single nucleotide variants (SNVs) at the *GYPA* locus using DNA resequencing data from a sample of populations from sub-Saharan Africa, Europe, East Asia, South Asia, Oceania, and the Americas. Tests for departures from neutrality were performed and the results from populations living in geographic regions of endemic *P. falciparum* (South Asia and sub-Saharan Africa) were compared to those from populations indigenous to non-endemic zones (East Asia, Europe, Oceania, and the Americas). Observed patterns of variation suggest that *GYPA* has been subject to balancing selection in populations living in malaria endemic areas. Of populations living in malaria non-endemic areas, no such evidence was found in samples from East Asia, Oceania, and the Americas but was found in samples from Europeans. While these results are consistent with malaria acting as a selective pressure on *GYPA*, they also suggest that another selective force has resulted in a similar pattern of variation in Europeans. Accordingly, *GYPA* has perhaps a more complex evolutionary history wherein on a global scale, spatially varying selective pressures have governed its natural history.

Sequence Variation across *GYPA*

We characterized the pattern of nucleotide diversity across *GYPA* in 85 individuals who are members of populations residing in regions where *P. falciparum* is endemic, sub-Saharan Africa and South Asia, and regions where it is non-endemic, East Asia and Europe (Fig. 1 and Table S1). Sequence data were obtained for each exon and ~1 kb upstream and downstream of each exon. Across all four populations, 37 variable sites were identified including five insertion/deletion (indel) polymorphisms and thirty-two single SNVs. The majority of these variants were intronic. Five SNVs, four of which were located in exon two (Fig. 2), were predicted to cause non-synonymous amino-acid substitutions. A single synonymous SNV was identified in exon two. Additionally, two SNVs were identified in the 5' UTR and one SNV was found in the 3' UTR of *GYPA*. Both of the SNVs that distinguish the MN blood groups were present, and MN allele frequencies and genotypes were consistent with published results (Table S2) (Mourant 1954; Race 1975).

If *GYPA* was subject to balancing selection from *P. falciparum*, we may well expect to observe relatively high levels of nucleotide diversity combined with departures from neutral expectations in populations living in regions where *P. falciparum* was endemic. Accordingly, we tested if patterns of nucleotide diversity and results of neutrality tests differed for populations living in malarial endemic zones compared to populations from non-endemic zones. Nucleotide diversity ranged from $\pi = 0.128$ in South Asians to $\pi = 0.17$ in sub-Saharan Africans with π estimates for European and East Asian populations falling in between, $\pi = 0.13$ and $\pi = 0.15$, respectively (Table 1). To place our nucleotide diversity estimates for *GYPA* in a genome-wide context, we compared our estimates to estimates obtained from the Environmental Genome Project's (EGP) exome sequencing project (NIEHS [National Institute of Environmental Health Sciences] <http://>

egp.gs.washington.edu). The EGP exome data consisted of genotypes from all protein-coding regions in the human genome sequenced in samples from 95 individuals including African Americans, sub-Saharan Africans (Yoruba), East Asians (Han Chinese from Beijing and Japanese from Tokyo), Latinos (Hispanic Americans), and European Americans (CEPH Europeans of North and West European Ancestry). No persons of South Asian ancestry were available for analysis. Compared to EGP data, our estimates of π were relatively low in sub-Saharan Africans (< 5%), Europeans (< 5%), and E. Asians (< 10%). Collectively, nucleotide variation in *GYP A* was lower than in more than ninety percent of known genes. Thus, if balancing selection acted on *GYP A*, it has not manifested as an excess of nucleotide diversity.

Next, we calculated F_{ST} between continental groups. F_{ST} is a measure of population divergence characterizing the degree of genetic differentiation between two or more populations. Estimates of F_{ST} that are higher than expected under expectations of neutrality may be the result of geographically restricted natural selection in one or more populations. Estimates of F_{ST} differed significantly ($p < 0.5$) between populations from sub-Saharan Africa and all other populations, with the F_{ST} between populations from sub-Saharan Africa and East Asia being the lowest (Table 2). No pairwise F_{ST} estimate between South Asians and Europeans and East Asians reached statistical significance (Table 2).

To place our F_{ST} estimates for *GYP A* in a genome-wide context, we estimated F_{ST} for all genes ($n = 16,911$) and a subset of genes that encode extracellular proteins ($n = 1,790$), assessed as part of the EGP. Pairwise *GYP A* F_{ST} estimates between sub-Saharan Africans and either Europeans or East Asians were in the top 10% ($F_{ST} > 0.091$) and top 15% ($F_{ST} > 0.010$) of the corresponding EGP distributions of F_{ST} , respectively. Results were similar when the analysis was limited to genes encoding extracellular domains. Accordingly, estimates of F_{ST} suggest that *P. falciparum* alone did not drive differentiation of *GYP A* in both populations from sub-Saharan Africa and South Asia, that selective pressure or duration of selection differed between populations in these two malaria endemic regions of the world, or some combination thereof.

Tajima's D can be used to detect departures from neutral expectations. Positive values of D can result from over-representation of intermediate frequency variants brought about by evolutionary forces (e.g., balancing selection) or demographic history (e.g., population structure). Negative values can reflect an excess of singletons, consistent with either directional selection or population expansion following a bottleneck. Tajima's D estimates for the *GYP A* locus were positive for sub-Saharan African, European, and South Asian populations and negative for East Asians (Table 1). None of the estimates of Tajima's D were significantly different from the neutral expectation based on a simulated distribution generated using the coalescent (Schaffner et al. 2005) or when compared to the empirical distribution of D values estimated from EGP exome data. Accordingly, we found no evidence that variation in *GYP A* differed from expectations under neutrality in the samples studied herein.

Sequence Variation in Exon 2 of *GYPA*

GYPA is located in a region with a high estimated recombination rate (Blumenfeld and Huang 1995). This potentially could reduce the sensitivity of tests that rely on the allele frequency spectra, such as Tajima's *D*, to detect departures from neutrality across *GYPA* (Thornton 2005). Furthermore, if *GYPA* has been evolving in response to malarial evasion, the effect might be strongest on exon 2, which includes the M and N alleles, and encodes the human extracellular domain that interacts directly with *P. falciparum*'s EBA-175 (Camus and Hadley 1985; Chitnis and Miller 1994; Sim et al. 1994; Tolia et al. 2005).

We tested whether the pattern of variation in exon 2 differed from other exons of *GYPA* in two ways. First, we calculated Tajima's *D* for exon 2 alone. Tajima's *D* for exon 2 was significantly positive ($p < 0.05$) in samples from sub-Saharan Africans, South Asians, and Europeans, but not East Asians (Table 1). Similar results were obtained when *D* values in sub-Saharan Africans ($p < 0.001$), Europeans ($p < 0.05$), and East Asians (N.S.) were compared to the empirical distribution generated using the EGP data. We also sequenced exon 2 in 949 Human Genome Diversity Panel-Centre d'Etude du Polymorphisme Humain (HGDP-CEPH) samples from 52 populations belonging to seven continental groups (Cann et al. 2002) and calculated Tajima's *D* values by continental group and by population. Tajima's *D* was significantly positive in sub-Saharan Africans ($D = 3.29$, $p < 0.01$) and Europeans ($D = 2.18$, $p < 0.05$) (Table 3). When analyzing Tajima's *D* by population, we observed a significant skew in the allele frequency spectra towards an excess of intermediate variants for six out of the seven sub-Saharan African populations. Only the San did not show a departure from neutrality ($D = 1.97$, NS) (Table S3). We also observed significant positive values of Tajima's *D* for six of the eight European populations (Table S3).

Next, we used a sliding windows approach to test if departures from neutral expectations varied along the length of *GYPA* and to identify the putative selected region(s). Two of the polymorphisms included in this analysis were indel polymorphisms, each of which was two base pairs (bp) in length. For these sites, each base pair was treated separately. Using a window size of 300 bp with an offset of 150 bp, 28 windows were tested across the 4,038 bp of sequence data. Among sub-Saharan African and South Asian populations, Tajima's *D* values were significantly positive ($p < 0.05$, based on 10,000 permutations) for windows four and five (Fig. 3 and Table S4). Estimates of Tajima's *D* were significantly positive for multiple windows across *GYPA* in samples from Europeans (Fig. 3). Collectively, these results are consistent with the hypothesis that malaria has driven the evolution of *GYPA* among sub-Saharan Africans and South Asians. It also suggests that a selective force other than malaria has shaped the evolution of *GYPA* in Europeans.

Haplotype Variation in *GYPA*

A network of haplotypes separated by mutational steps in a region that has been influenced by balanced selection can have a "dumbbell-like" pattern wherein two clusters of closely related haplotypes are separated from one another by multiple mutation steps. We constructed a reduced median network illustrating the relationships among a total of 34 identified *GYPA* haplotypes (Fig. 4 and Table S5). Five *GYPA* haplotypes were common (frequency = 5%) and twenty-nine haplotypes were rare (frequency < 5%). Sixteen

haplotypes occurred only once in the sample, and seven of these were geographically restricted to sub-Saharan Africa. The two most common haplotypes, H1 and H2, were found in all four populations and account for 48% of sampled chromosomes. H1 contained two SNVs that distinguish the haplotype containing the N allele whereas H2 contained the two SNVs that define the haplotype containing the M allele. Consistent with these observations, a network of *GYPA* haplotypes revealed two clusters of common haplotypes connected by a single mutational step to multiple rare haplotypes and separated from one another by multiple mutational steps. Color-coding the haplotypes according to whether they carried the M or N alleles highlighted this pattern. These results are consistent with balancing selection maintaining the M and N alleles of *GYPA* in multiple populations and not limited to sub-Saharan Africans and South Asians, and reinforce the suggestion that selective forces other than *falciparum* malaria have acted upon *GYPA*.

Discussion

GYPA is a cell-surface glycoprotein that serves as the binding ligand for *P. falciparum* erythrocyte invasion. It stands out among loci in the human genome as one of the fastest evolving genes. Our results suggest that *GYPA* evolved under balancing selection among two populations from *P. falciparum* endemic zones, sub-Saharan Africans and South Asians, and one population from a *P. falciparum* non-endemic zone, Europeans. The evidence for balancing selection both within and outside of malarial endemic zones could be explained in two ways: 1) spatially varying selection pressures in response to localized environmental variables have shaped *GYPA* evolution on a global scale or 2) a single shared selective pressure common to Europeans, sub-Saharan Africans, and South Asians. Both of these interpretations point to a potential, as of yet unidentified, role of GPA in infectious disease pathways beyond that of *P. falciparum*.

Evidence for spatially varying selection pressure is supported by three observations. First, two alleles, M and N, have been maintained at high frequency in all three populations. Second, among sub-Saharan Africans and South Asians we observe positive and significant values of Tajima's *D* that are localized to exon two, which encodes the extracellular domain that interacts directly with *P. falciparum*. Third, among Europeans we observe significantly positive values of Tajima's *D* across multiple protein domains and two alleles separated by multiple mutational steps.

The skew in the allele frequency spectra towards an excess of intermediate frequency variants was observed in sub-Saharan Africans and South Asians and was restricted to exon 2. *GYPA* exon 2 encodes the extracellular protein domain that interacts directly with *P. falciparum*'s EBA-175 ligand leading to merozoite invasion of human erythrocytes. Haplotype networks for exon 2 in South Asians and sub-Saharan Africans showed two major haplotypes corresponding to the M and N of the MNS blood groups encoded by two non-synonymous variants in exon 2. The maintenance of these two alleles at intermediate frequencies and occurring on separate haplotypes separated by multiple mutational steps supports the hypothesis that this locus has been evolving under balancing selection to *P. falciparum*.

Population subdivision also can lead to significant departures from neutrality and positive values of Tajima's D . However, we do not think the pattern of variation in exon 2 results from underlying demographic factors for several reasons. First, demographic events are evolutionary forces that affect different genomic regions in a similar manner, whereas natural selection is a locus-specific force causing deviations from the expected neutral variation in both functional variants and markers closely linked to them. We observe a departure from neutrality that is restricted to exon 2. Second, we accounted for underlying demography and mitigated its effect by using simulated and empirical distributions to test for statistical significance in each continental population. Third, we detected significant deviations across multiple sub-Saharan populations with different demographic histories. Together, these observations suggest that natural selection, not demography, is responsible for the observed pattern of variation among sub-Saharan Africans and South Asians.

We acknowledge that the statistical tests used are not independent of one another. Unfortunately, there is no straightforward way to correct for this lack of independence given the various statistical tests employed and the numerous population combinations. This is a limitation of our study design.

Balancing selection acts to maintain diversity at a locus most commonly via heterozygote advantage or frequency dependent selection. Many of the mutations that confer resistance to malaria exhibit heterozygote advantage and are deleterious outside of malarial endemic zones (i.e. HbS allele and *G6PD* A- allele) (Tishkoff et al. 2001). Under a scenario of heterozygote advantage for *GYPA*, one prediction is that persons who are MN heterozygotes might be more resistant to *P. falciparum* infection and/or disease progression than individuals homozygous for either allele. However, it remains uncertain if MN heterozygosity provides resistance to malaria.

Under frequency dependent selection, the advantageous allele would switch back and forth between the M and N alleles and be dependent on the frequency of a corresponding allele encoding *P. falciparum*'s EBA-175 binding antigen. The pattern of variation in the EBA-175 binding domain suggests that selection has maintained amino acid polymorphism in this region (Ozwar et al. 2001). However, the specific alleles facilitating the parasite-erythrocyte interaction remain unknown. Therefore, it remains to be determined if the observed patterns in humans and malaria are the result of frequency dependent selection. Continued research into the relationship between diversity patterns of *GYPA*'s extracellular domain and *P. falciparum*'s EBA-175 antigen will be necessary to determine the type of balancing selection to malarial pressure operating on this locus.

Analysis of exon 2 suggests that balancing selection acted on populations outside of hyperendemic or mesoendemic malarial zones. Specifically, Tajima's D for exon 2 was significant and positive for Europeans and for six of eight European HGDP-CEPH populations tested separately. Moreover, our sliding-window analysis revealed evidence that balancing selection acted on genomic regions beyond the boundaries of exon 2. This observation indicates that the target of balancing selection resides outside of the region encoding the extracellular domain that interacts with *P. falciparum* suggesting that a selective force other than, or in combination with, *P. falciparum* is responsible. In other

words, these observations suggest that *GYPA* has been subject to spatially varying selection wherein distinct selective pressures have operated in particular environments.

It is challenging to speculate what pathogen, if any, could be driving the pattern of selection observed for Europeans. There are no known pathogens present in Europe that infect erythrocytes and bind to GPA. However, erythrocyte glycoproteins including GPA do bind many viral and microbial pathogens that do not infect the red cell. This observation has been used to develop the hypothesis that glycoproteins serve as decoy receptors. They attract pathogens to enucleated erythrocytes and divert pathogens away from target tissues (Baum et al. 2002; Gagneux and Varki 1999). It is possible that in the absence of *falciparum* malaria, allelic variation has been maintained to attract diverse pathogens. Alternatively, a single shared selective pressure common to Europeans, sub-Saharan Africans, and South Asians could result in the pattern of balancing selection observed. However, there are no known pathogens common to all three regions that bind to GPA and infect the erythrocyte. *P. falciparum* is the only known pathogen that both binds to GPA and infects erythrocytes. Other pathogens bind to GYPA, but do not infect red cells. These include viruses such as rotavirus (Salas et al. 2016), hepatitis A (Sanchez et al. 2004), and influenza (Burness and Pardoe 1981) as well as bacteria including *Mycoplasma pneumoniae* (Baseman et al. 1982) and *Escherichia coli* (Brooks et al. 1989).

Conclusion

Malaria has been a powerful selective force in human evolution. It has shaped the patterns of human genetic variation at multiple loci in the human genome including beta-globin and *G6PD* (Hamblin and Di Rienzo 2000; Saunders et al. 2002; Tishkoff et al. 2001). The most virulent form of malaria, *P. falciparum*, invades the human host by docking to GPA expressed on the erythrocyte surface, specifically the extracellular domain encoded by exon 2 of the *GYPA* gene. We identified an excess of intermediate frequency variants in exon 2 of *GYPA* in three continental groups; sub-Saharan Africans, South Asians, and Europeans. Two of these groups (sub-Saharan Africans and South Asians) inhabit regions of the globe with endemic *P. falciparum*, but Europeans do not. Our results suggest that balancing selection has acted on exon 2 of *GYPA* in *P. falciparum* endemic regions, but that a selective force other than *P. falciparum* has acted on *GYPA* in European populations. This conclusion is not entirely surprising and indeed, for genes encoding proteins that interact with a multitude of environmental factors, the pattern of selection might be expected to be complex and characterized by the impact of different selective forces, with varied strength, and acting at different periods of human evolutionary history.

Materials and Methods

Populations

GYPA was sequenced in 85 individuals (i.e., the discovery panel) including 26 sub-Saharan Africans, 17 Europeans, 14 East Asians, and 28 South Asians (Fig. 1 and Table S1). The sub-Saharan Africans consisted of samples from ten ethnic groups (2 Biaka, 3 Alur, 3 Hema, 3 Nande, 3 Mbuti Pygmy, 1 Kenyan, 3 Nigerians, 2 San, 3 Sotho, and 3 Nguni). The South Asian samples were from ten populations (3 Brahmin, 3 Chenchu, 3 Irula, 3 Khondadora, 3

Kshatriya, 3 Madiga, 3 Mala, 3 Relli, 1 Santal, and 3 Yadava). East Asians were represented by samples from five populations (3 Han Chinese, 3 Cambodians, 2 Vietnamese, 3 Japanese, and 3 Aboriginal Malay) and Europeans by samples from seven populations (3 French, 3 Northern Europeans, 3 Poles, 4 Daghestani, 1 Italian, 1 Finn, and 2 Turku). To increase the breadth of populations studied and samples tested, we sequenced exon 2 of *GYP A* in 954 individuals from 52 populations from sub-Saharan Africa, the Middle East, Europe, East Asia, South/Central Asia, the Americas, and Oceania (Table S6) that are part of the Human Genome Diversity Panel-Centre d'Etude du Polymorphisme Humain (HGDP-CEPH) (Cann et al. 2002). Duplicate samples and first-degree relatives were removed (Rosenberg 2006).

DNA Sequencing

A set of seven primer pairs was used to amplify a total of 4 kb of sequence including each exon and approximately 1 kb of sequence upstream and downstream of each exon. The HGDP-CEPH samples were sequenced only for the region containing exon 2. Sequence data were generated using internal primers on either an ABI 3700 or ABI 3730xl automated sequencer. PCR and sequencing primers are provided in Table S7. Sequence trace files were aligned and genotype calls were made using Codon Code (Dedham, MA). Polymorphic sites were confirmed by visual inspection of the traces. Genotype data is available in the Supplementary data online.

Statistical Analysis

Pairwise and population specific F_{ST} were calculated using Arlequin version 3.1 (Excoffier et al. 2005). Tajima's D detects departures from neutrality based on the allele frequency spectra and was calculated using DnaSP Version 5.0 (<http://www.ub.edu/dnasp/>) (Librado and Rozas 2009; Tajima 1989). Statistical significance of each test was assessed using the coalescent as implemented in the program COSI (Schaffner et al. 2005). This program incorporates population growth into the modeling procedure, the inclusion of which more accurately reflects the demographic history of humans. We used 10,000 replicates and modeled the assumed demographic history of each of our four populations. For Africans, East Asians, and Europeans, the default parameters were implemented. For South Asians, we selected parameters that reflected what is known about South Asian demographic history. The parameter file used in our COSI analysis is provided in Supplemental Data File S1. All exons of *GYP A* and exons encoding just the extracellular domain were analyzed in each population. Genotypes were phased using fastPhase (Scheet and Stephens 2006). Haplotype networks were constructed using Network v4.5.1.0 ([Fluxus-Engineering.com](http://fluxus-engineering.com)) (Bandelt et al. 1995; Polzin and Daneshmand 2003).

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

AWB was supported by a training fellowship from the NIH–National Human Genome Research Institute (T32HG00035). The authors thank Anita Beck, Kati Buckingham, and Heidi Gildersleeve for thoughtful discussions on the manuscript.

References

- Bamshad M, Wooding SP (2003) Signatures of natural selection in the human genome. *Nature reviews. Genetics* 4: 99–111. doi: 10.1038/nrg999
- Bandelt HJ, Forster P, Sykes BC, Richards MB (1995) Mitochondrial portraits of human populations using median networks. *Genetics* 141: 743–53. [PubMed: 8647407]
- Barreiro LB, Quintana-Murci L (2010) From evolutionary genetics to human immunology: how selection shapes host defence genes. *Nature reviews. Genetics* 11: 17–30. doi: 10.1038/nrg2698
- Baseman JB, Banai M, Kahane I (1982) Sialic acid residues mediate *Mycoplasma pneumoniae* attachment to human and sheep erythrocytes. *Infection and immunity* 38: 389–91. [PubMed: 6815092]
- Baum J, Ward RH, Conway DJ (2002) Natural selection on the erythrocyte surface. *Molecular biology and evolution* 19: 223–9. [PubMed: 11861881]
- Blumenfeld OO, Huang CH (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. *Human mutation* 6: 199–209. doi: 10.1002/humu.1380060302 [PubMed: 8535438]
- Brooks DE, Cavanagh J, Jayroe D, Janzen J, Snoek R, Trust TJ (1989) Involvement of the MN blood group antigen in shear-enhanced hemagglutination induced by the *Escherichia coli* F41 adhesin. *Infection and immunity* 57: 377–83. [PubMed: 2563256]
- Burness AT, Pardoe IU (1981) Effect of enzymes on the attachment of influenza and encephalomyocarditis viruses to erythrocytes. *J Gen Virol* 55: 275–88. doi: 10.1099/0022-1317-55-2-275 [PubMed: 6270264]
- Camus D, Hadley TJ (1985) A *Plasmodium falciparum* antigen that binds to host erythrocytes and merozoites. *Science* 230: 553–6. [PubMed: 3901257]
- Chitnis CE, Miller LH (1994) Identification of the erythrocyte binding domains of *Plasmodium vivax* and *Plasmodium knowlesi* proteins involved in erythrocyte invasion. *The Journal of experimental medicine* 180: 497–506. [PubMed: 8046329]
- Excoffier L, Laval G, Schneider S (2005) Arlequin (version 3.0): an integrated software package for population genetics data analysis. *Evolutionary bioinformatics online* 1: 47–50.
- Gagneux P, Varki A (1999) Evolutionary considerations in relating oligosaccharide diversity to biological function. *Glycobiology* 9: 747–55. [PubMed: 10406840]
- Hamblin MT, Di Rienzo A (2000) Detection of the signature of natural selection in humans: evidence from the Duffy blood group locus. *Am J Hum Genet* 66: 1669–79. [PubMed: 10762551]
- Karlsson KA (1995) Microbial recognition of target-cell glycoconjugates. *Current opinion in structural biology* 5: 622–35. [PubMed: 8574698]
- Ko WY, Kaercher KA, Giombini E, Marcatali P, Froment A, Ibrahim M, Lema G, Nyambo TB, Omar SA, Wambebe C, Ranciaro A, Hirbo JB, Tishkoff SA (2011) Effects of natural selection and gene conversion on the evolution of human glycophorins coding for MNS blood polymorphisms in malaria-endemic African populations. *American journal of human genetics* 88: 741–54. doi: 10.1016/j.ajhg.2011.05.005 [PubMed: 21664997]
- Librado P, Rozas J (2009) DnaSP v5: a software for comprehensive analysis of DNA polymorphism data. *Bioinformatics* 25: 1451–2. doi: 10.1093/bioinformatics/btp187 [PubMed: 19346325]
- Mourant AE (1954) *The Distribution of the Human Blood Groups*. Blackwell Scientific Publications, Oxford
- Nielsen R, Hellmann I, Hubisz M, Bustamante C, Clark AG (2007) Recent and ongoing selection in the human genome. *Nature reviews. Genetics* 8: 857–68. doi: 10.1038/nrg2187
- Nishimura H, Sugawara K, Kitame F, Nakamura K (1988) Attachment of influenza C virus to human erythrocytes. *The Journal of general virology* 69 (Pt 10): 2545–53. [PubMed: 3049938]
- Okoyeh JN, Pillai CR, Chitnis CE (1999) *Plasmodium falciparum* field isolates commonly use erythrocyte invasion pathways that are independent of sialic acid residues of glycophorin A. *Infection and immunity* 67: 5784–91. [PubMed: 10531229]

- Onda M, Kudo S, Fukuda M (1994) Genomic organization of glycophorin A gene family revealed by yeast artificial chromosomes containing human genomic DNA. *The Journal of biological chemistry* 269: 13013–20. [PubMed: 8175721]
- Ozwarra H, Kocken CH, Conway DJ, Mwenda JM, Thomas AW (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. *Molecular and biochemical parasitology* 116: 81–4. [PubMed: 11463469]
- Pasvol G, Wainscoat JS, Weatherall DJ (1982) Erythrocytes deficiency in glycophorin resist invasion by the malarial parasite *Plasmodium falciparum*. *Nature* 297: 64–6. [PubMed: 7040988]
- Paul RW, Lee PW. (1987) Glycophorin is the reovirus receptor on human erythrocytes. *Virology* 159: 94–101. [PubMed: 3604060]
- Polzin T, Daneshmand SV (2003) On Steiner trees and minimum spanning trees in hypergraphs. *Operations Research Letters* 31: 12–20.
- Race RR, Sanger R. (1975) *The MNS blood groups*. . 6 edn Blackwell, Oxford, pp 92–138
- Rearden A, Magnet A, Kudo S, Fukuda M (1993) Glycophorin B and glycophorin E genes arose from the glycophorin A ancestral gene via two duplications during primate evolution. *The Journal of biological chemistry* 268: 2260–7. [PubMed: 8420995]
- Saada AB, Terespolski Y, Adoni A, Kahane I (1991) Adherence of *Ureaplasma urealyticum* to human erythrocytes. *Infection and immunity* 59: 467–9. [PubMed: 1987063]
- Salas A, Marco-Puche G, Trivino JC, Gomez-Carballa A, Cebeay-Lopez M, Rivero-Calle I, Vilanova-Trillo L, Rodriguez-Tenreiro C, Gomez-Rial J, Martinon-Torres F (2016) Strong down-regulation of glycophorin genes: A host defense mechanism against rotavirus infection. *Infect Genet Evol* 44: 403–11. doi: 10.1016/j.meegid.2016.07.044 [PubMed: 27491455]
- Sanchez G, Aragonés L, Costafreda MI, Ribes E, Bosch A, Pinto RM (2004) Capsid region involved in hepatitis A virus binding to glycophorin A of the erythrocyte membrane. *J Virol* 78: 9807–13. doi: 10.1128/JVI.78.18.9807-9813.2004 [PubMed: 15331714]
- Saunders MA, Hammer MF, Nachman MW (2002) Nucleotide variability at G6pd and the signature of malarial selection in humans. *Genetics* 162: 1849–61. [PubMed: 12524354]
- Schaffner SF, Foo C, Gabriel S, Reich D, Daly MJ, Altshuler D (2005) Calibrating a coalescent simulation of human genome sequence variation. *Genome research* 15: 1576–83. doi: 10.1101/gr.3709305 [PubMed: 16251467]
- Scheet P, Stephens M (2006) A fast and flexible statistical model for large-scale population genotype data: Applications to inferring missing genotypes and haplotypic phase. *American journal of human genetics* 78: 629–644. [PubMed: 16532393]
- Sim BK, Chitnis CE, Wasniowska K, Hadley TJ, Miller LH (1994) Receptor and ligand domains for invasion of erythrocytes by *Plasmodium falciparum*. *Science* 264: 1941–4. [PubMed: 8009226]
- Tajima F (1989) Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics* 123: 585–95. [PubMed: 2513255]
- Tavakkol A, Burness AT (1990) Evidence for a direct role for sialic acid in the attachment of encephalomyocarditis virus to human erythrocytes. *Biochemistry* 29: 10684–90. [PubMed: 2176879]
- Thornton K (2005) Recombination and the properties of Tajima's D in the context of approximate-likelihood calculation. *Genetics* 171: 2143–8. doi: 10.1534/genetics.105.043786 [PubMed: 15998723]
- Tishkoff SA, Varkonyi R, Cahinhinan N, Abbes S, Argyropoulos G, Destro-Bisol G, Drousiotou A, Dangerfield B, Lefranc G, Loiselet J, Piro A, Stoneking M, Tagarelli A, Tagarelli G, Touma EH, Williams SM, Clark AG (2001) Haplotype diversity and linkage disequilibrium at human G6PD: recent origin of alleles that confer malarial resistance. *Science* 293: 455–62. doi: 10.1126/science.1061573 [PubMed: 11423617]
- Tolia NH, Enemark EJ, Sim BK, Joshua-Tor L (2005) Structural basis for the EBA-175 erythrocyte invasion pathway of the malaria parasite *Plasmodium falciparum*. *Cell* 122: 183–93. doi: 10.1016/j.cell.2005.05.033 [PubMed: 16051144]
- Verrelli BC, McDonald JH, Argyropoulos G, Destro-Bisol G, Froment A, Drousiotou A, Lefranc G, Helal AN, Loiselet J, Tishkoff SA (2002) Evidence for balancing selection from nucleotide

sequence analyses of human G6PD. American journal of human genetics 71: 1112–28. doi: 10.1086/344345 [PubMed: 12378426]

Wang HY, Tang H, Shen CK, Wu CI (2003) Rapidly evolving genes in human. I. The glycophorins and their possible role in evading malaria parasites. Molecular biology and evolution 20: 1795–804. doi: 10.1093/molbev/msg185 [PubMed: 12949139]

Wood ET, Stover DA, Slatkin M, Nachman MW, Hammer MF (2005) The beta -globin recombinational hotspot reduces the effects of strong selection around HbC, a recently arisen mutation providing resistance to malaria. American journal of human genetics 77: 637–42. doi: 10.1086/491748 [PubMed: 16175509]

Wybenga LE, Epand RF, Nir S, Chu JW, Sharom FJ, Flanagan TD, Epand RM (1996) Glycophorin as a receptor for Sendai virus. Biochemistry 35: 9513–8. doi: 10.1021/bi9606152 [PubMed: 8755731]

Fig. 1.

Geographic distribution of the resequencing panel populations. The European sample (blue) consists of 17 individuals from 7 populations including: 3 French, 3 Polish, 4 Daghestani, 1 Italian, and 3 Finnish, and 3 Northern Europeans of unknown ethnic origin. The sub-Saharan African samples (red) consist of individuals from 10 ethnic groups including: 2 Biaka, 3 Alur, 3 Hema, 3 Nande, 3 Mbuti, 1 Kenyan, 3 Nigerians, 2 San, 3 Sotho, and 3 Nguni. The South Asian samples (turquoise) are comprised of 28 individuals from 10 populations including: 3 Brahmin, 3 Chenchu, 3 Irula, 3 Khondadora, 3 Kshatriya, 3 Madiga, 3 Mala, 3 Relli, 1 Santal, and 3 Yadava. The samples from East Asians (green) are from 5 ethnic groups including: 3 Han Chinese, 3 Cambodians, 2 Vietnamese, 3 Japanese, and 3 Aboriginal Malay.

Fig. 2.

Variants identified in glycophorin A (*GYPA*). Schematic representation of the intron/exon structure and the 43 SNVs identified in the resequencing panel and the HGDP-CEPH panel. Each exon is colored according to the cellular domain it encodes. The signal peptide is purple, the extracellular domain is blue, the transmembrane domain is green, and the intracellular domain is yellow. Non-synonymous changes are indicated with an asterisk.

Fig. 3. Estimates of Tajima's D (A) from windows 300 bp in length with 150 bp offsets across *GYP A*. Exons of *GYP A* are depicted as filled squares above the X-axis. The test statistics peak in sub-Saharan African and South Asian populations in the windows encompassing exon 2.

Fig. 4. Reduced median network of *GYPA* haplotypes from sub-Saharan Africans (yellow), East Asians (green), Europeans (red), and South Asians (blue). Each circle represents a unique haplotype and the size of each circle corresponds to its frequency. Each grey circle represents a single inferred mutational step. The inset shows the same network typology, but with color-coding according to MN genotype, M (black) and N (orange).

Table 1.

Summary statistics calculated for the length of *GYPA* and for each exon encoding the extracellular domain the four populations comprising the discovery panel.

Gene Region	Population	<i>n</i>	Nucleotide Sites	<i>S</i>	π	$\theta\omega$	Tajima's D^a
GYPA	sub-Saharan African	25	31	21	0.17	4.91	0.24
	East Asian	13	26	16	0.15	4.19	-0.15
	European	16	29	11	0.13	2.73	1.29
	South Asian	27	26	11	0.13	2.41	1.08
	Endemic	52	26	21	0.16	4.22	-0.10
	Non-Endemic	29	26	19	0.14	4.11	-0.27
Exon 2	sub-Saharan African	25	5	5	0.43	1.12	2.20**
	East Asian	13	5	5	0.30	1.31	0.37
	European	17	5	3	0.27	0.73	1.92**
	South Asian	28	5	3	0.30	0.65	2.55**
	Endemic	53	5	5	0.36	0.96	1.89**
	Non-Endemic	30	5	5	0.28	1.07	0.69*
Exon 3	sub-Saharan African	26	0	0	NA	NA	NA
	East Asian	14	0	0	NA	NA	NA
	European	17	0	0	NA	NA	NA
	South Asian	28	0	0	NA	NA	NA
	Endemic	54	0	0	NA	NA	NA
	Non-Endemic	31	0	0	NA	NA	NA
Exon 4	sub-Saharan African	26	0	0	NA	NA	NA
	East Asian	14	0	0	NA	NA	NA
	European	17	0	0	NA	NA	NA
	South Asian	28	0	0	NA	NA	NA
	Endemic	54	0	0	NA	NA	NA
	Non-Endemic	31	0	0	NA	NA	NA

Statistical significance was assessed using the theoretical distribution.

n = number of individuals; *S* = segregating sites; π = average number of nucleotide differences per site; $\theta\omega$ = Watterson estimator;

**
p<0.01;

*
p<0.05

^aTajima (1989)

Table 2.

Pairwise F_{ST} estimates between the four discovery-panel populations.

	sub-Saharan African	East Asian	South Asian	European
sub-Saharan African		p<0.05	p<0.05	p<0.05
East Asian	0.087		N.S.	N.S.
South Asian	0.099	-0.005		N.S.
European	0.111	-0.023	-0.012	

Pairwise F_{ST} is reported below the diagonal line and statistical significance is reported above the diagonal line.

Table 3.

Summary statistics among the HGDP-CEPH continental groups.

Region	population	n	S	π	$\theta\omega$	Tajima's D^a
Exon 2 region	East Asian	478	8	0.15	1.19	1.50
	European	312	6	0.15	0.95	2.40 [*]
	Indigenous American	58	7	0.16	1.51	1.26
	Middle Eastern	324	7	0.16	1.10	2.19 [*]
	Oceanian	60	6	0.13	1.29	0.92
	Central Asian	388	9	0.14	1.38	0.89
	sub-Saharan African	210	10	0.18	1.69	1.13
Exon 2	East Asian	478	5	0.31	0.74	1.85
	European	312	4	0.29	0.63	2.18 [*]
	Indigenous American	58	5	0.35	1.08	1.46
	Middle Eastern	324	5	0.32	0.79	1.90
	Oceanian	60	5	0.31	1.07	1.00
	Central Asian	388	5	0.28	0.76	1.44
	sub-Saharan African	210	5	0.46	0.84	3.29 ^{**}

Exon 2 includes only variants present in the exon whereas exon 2 region includes intronic variants present upstream and downstream of exon 2.

Statistical significance was assessed using 10,000 permutation tests in DnaSP (Librado and Rozas 2009).

n = number of individuals; S = segregating sites; π = average number of nucleotide differences per site; $\theta\omega$ = Watterson estimator;

^{**} p < 0.001;

^{*} p < 0.01; ^{*}p < 0.05

^a Tajima (1989)