Widespread balancing selection and pathogen-driven selection at blood group antigen genes

Matteo Fumagalli,[1,2] Rachele Cagliani,[1] Uberto Pozzoli,[1] Stefania Riva,[1] Giacomo P. Comi,[3] Giorgia Menozzi,[1] Nereo Bresolin,[1,3] and Manuela Sironi[1,4]

[1] *Scientific Institute IRCCS E. Medea, Bioinformatic Lab, 23842 Bosisio Parini (LC), Italy;* [2] *Bioengineering Department, Politecnico di Milano, 20133 Milan, Italy;* [3] *Dino Ferrari Centre, Department of Neurological Sciences, University of Milan, IRCCS Ospedale Maggiore Policlinico, Mangiagalli and Regina Elena Foundation, 20100 Milan, Italy*

Historically, allelic variations in blood group antigen (BGA) genes have been regarded as possible susceptibility factors for infectious diseases. Since host–pathogen interactions are major determinants in evolution, BGAs can be thought of as selection targets. In order to verify this hypothesis, we obtained an estimate of pathogen richness for geographic locations corresponding to 52 populations distributed worldwide; after correction for multiple tests and for variables different from selective forces, significant correlations with pathogen richness were obtained for multiple variants at 11 BGA loci out of 26. In line with this finding, we demonstrate that three BGA genes, namely *CD55*, *CD151*, and *SLC14A1*, have been subjected to balancing selection, a process, rare outside MHC genes, which maintains variability at a locus. Moreover, we identified a gene region immediately upstream the transcription start site of *FUT2* which has undergone non-neutral evolution independently from the coding region. Finally, in the case of *BSG*, we describe the presence of a highly divergent haplotype clade and the possible reasons for its maintenance, including frequency-dependent balancing selection, are discussed. These data indicate that BGAs have been playing a central role in the host–pathogen arms race during human evolutionary history and no other gene category shows similar levels of widespread selection, with the only exception of loci involved in antigen recognition.

[Supplemental material is available online at www.genome.org.]

Since the discovery of the ABO blood group in 1900 by Karl Landsteiner, as many as 29 blood group (BG) systems have been identified in humans (Blood Group Antigen Gene Mutation Database, BGMUT [Blumenfeld and Patnaik 2004]). Each system is specified by a blood group antigen (BGA) constituted by a protein or carbohydrate molecule which is expressed on the erythrocyte membrane and is polymorphic in human populations.

The molecular basis of all blood group systems (except for P1) has been clarified, with one or more polymorphic loci accounting for BG phenotypes. BGA genes belong to different functional categories, including receptors, transporters, channels, adhesion molecules, and enzymes; among the latter, the great majority of loci code for glycotransferases. While a few BGAs are confined to the erythrocyte membrane, others are expressed at the surface of different cell types or secreted in body fluids (Reid and Lomas-Frances 1997).

The number of different alleles is highly variable among BGA genes and ranges from two to >100 (Blumenfeld and Patnaik 2004) with the most common form of variation being accounted for by missense or nonsense single nucleotide polymorphisms (SNPs). BGA polymorphisms have attracted considerable attention over recent years not only with respect to erythrocyte physiology per se, but also due to the possibility that variations in BGAs might underlie different susceptibility to diseases. In particular, the association between infections and BGA polymorphisms has been extensively investigated, although conclusive results have been obtained in a minority of cases. For example, specific BGA alleles have been shown to alter susceptibility to malaria (Moulds and Moulds 2000), while *FUT2* variants (Lewis system) influence the predisposition to Norwalk virus (Lindesmith et al. 2003) and *Campylobacter* (Ruiz-Palacios et al. 2003) infection, as well as to vulvovaginal candidisis (Hurd and Domino 2004) and urinary tract infections (Schaeffer et al. 2001).

Such findings are in line with the vision whereby different BGAs serve as "incidental receptors for viruses and bacteria" (Moulds et al. 1996), but also function as modulators of innate immune response (Ruiz-Palacios et al. 2003; Linden et al. 2008) and possibly as "decoy-sink" molecules targeting pathogens to macrophages (Gagneux and Varki 1999).

Given this premise and the conundrum whereby host–pathogen interactions are major determinants in evolution, BGAs can be thought of as possible targets of diverse selective pressures. This view is in agreement with the geographic differentiation pattern observed for BGAs and with previous reports of non-neutral evolution at the *ABO*, *DARC*, *GYPA*, and *FUT2* loci (Saitou and Yamamoto 1997; Koda et al. 2001; Baum et al. 2002; Hamblin et al. 2002; Calafell et al. 2008).

Here we exploited the availability of extensive resequencing data, as well as of SNP genotyping in world-wide populations, to investigate the evolutionary forces underlying the evolution of BGA genes: Our data provide evidence that balancing and pathogen-driven selections have acted at multiple BGA loci.

Results and Discussion

Pathogen richness and BGA gene polymorphisms

As a first step we wished to verify whether allele frequencies of SNPs in BGA genes varied with pathogen richness, in terms of

[4] **Corresponding author.**
E-mail manuela.sironi@bp.lnf.it; fax 39-031-877499.

Article published online before print. Article and publication date are at http://www.genome.org/cgi/doi/10.1101/gr.082768.108.

different species in a geographic location. Similar approaches have been applied to test this same hypothesis for HLA genes (Prugnolle et al. 2005) and for other gene–environment interactions (Thompson et al. 2004; Young et al. 2005; Hancock et al. 2008). To this aim we exploited the fact that a set of over 650,000 tag SNPs has been typed in 52 populations (HGDP-CEPH panel) distributed world wide (Li et al. 2008). As for pathogen richness, we gathered information concerning the number of different micropathogen species from the Gideon database; pathogen richness was calculated on a country basis by pooling together viruses, bacteria, fungi and protozoa (see Methods for further details). A total of 262 SNPs in BGA genes had been typed in the HGDP-CEPH panel allowing analysis of the following loci: *RHCE, ERMAP, DARC, CD55, CR1, GYPC, GYPA, GCNT2, RHAG, C1GALT1, AQP1, KEL, AQP3, ABO, CD44, ART4* (also known as *DO*), *SEMA7A, SLC4A1, SLC14A1, FUT3, BCAM* (also known as *LU*), *FUT2, FUT1, A4GALT, XG*, and *XK*.

For all 262 BGA SNPs in the data set we calculated Kendall's rank correlation coefficient (τ) between pathogen richness and allele frequencies in HGDP-CEPH populations; a normal approximation with continuity correction to account for ties was used for *P*-value calculations (Kendall 1976). We verified that, after Bonferroni correction for multiple tests, 26 BGA gene SNPs were significantly associated with pathogen richness (Table 1). Since variables different from selective forces (e.g., colonization routes; Handley et al. 2007) are expected to affect allele frequency spectra across populations, we compared the strength of BGA gene SNP correlations to control sets of SNPs extracted from the data set. In particular, for each BGA SNP in Table 1 we extracted from the full data set all SNPs having an overall minor allele frequency (averaged over all populations) differing less than 0.01 from its frequency; for all SNPs in the 26 frequency-matched groups we calculated Kendall's τ between pathogen richness and allele frequencies. Next, we calculated the percentile rank of BGA gene SNPs in the distribution of Kendall's τ obtained for the control sets and in the distribution of all SNPs in the data set. Data are reported in Table 1 and indicate that all SNPs ranked above the 90th percentile of τ-values, with 19 of them ranking above the 95th (data for all 262 SNPs are available in Supplemental File 1). By performing 30,000 simulations using samples of 262 SNPs we verified that the probability of obtaining 19 SNPs with a correlation value above the 95th percentile amounted to 0.045; a similar result is obtained by considering the probability to obtain *n* SNPs with a τ higher than the 95th percentile in a sample of 262 to be Poisson-distributed (P = 0.043). These data therefore indicate that the fraction of BGA SNPs that correlate with pathogen richness is higher than expected; yet, these calculations also suggest that a portion of SNPs in Table 1 might represent false positive associations in that the retrieval of 13 variants with a percentile rank above the 95th would be expected by chance. An estimation of the magnitude of selective effects exerted by pathogens on human genes would be required to accurately estimate the expected fraction of truly correlated SNPs.

Table 1. **BGA gene SNPs significantly associated with pathogen richness**

			P	Rank[b]	Rank[c]
SNP	Gene	τ[a]	(Bonferroni)	(all)	(matched)
rs11210729	*ERMAP*	0.446	0.002	0.945	0.937
rs6700168	*CD55*	0.427	0.005	0.923	0.927
rs4143022	*GYPC*	−0.440	0.003	0.938	0.930
rs7589096	*GYPC*	0.548	0.000	0.997	0.997
rs4663038	*GYPC*	−0.460	0.001	0.958	0.961
rs17741574	*GYPC*	0.459	0.002	0.957	0.953
rs13034269	*GYPC*	0.493	0.001	0.981	0.978
rs6568	*GYPC*	0.417	0.009	0.910	0.911
rs10487590	*C1GALT1*	0.549	0.000	0.997	0.997
rs9466910	*GCNT2*	−0.491	0.001	0.979	0.975
rs9466912	*GCNT2*	−0.491	0.001	0.979	0.975
rs17576994	*GCNT2*	0.476	0.001	0.970	0.972
rs2228332	*AQP3*	−0.459	0.001	0.957	0.957
rs2073824	*ABO*	−0.506	0.000	0.987	0.988
rs2421826	*CD44*	−0.436	0.004	0.933	0.929
rs1547059	*CD44*	0.439	0.006	0.937	0.933
rs2072081	*SLC4A1*	0.487	0.000	0.978	0.979
rs2074108	*SLC4A1*	0.473	0.001	0.968	0.970
rs692899	*SLC14A1*	0.425	0.007	0.920	0.916
rs10853535	*SLC14A1*	−0.509	0.000	0.988	0.985
rs566309	*SLC14A1*	0.461	0.005	0.958	0.951
rs900971	*SLC14A1*	−0.611	0.000	1.000	1.000
rs6507641	*SLC14A1*	−0.466	0.001	0.962	0.963
rs602662	*FUT2*	−0.499	0.000	0.984	0.980
rs485186	*FUT2*	−0.513	0.000	0.989	0.987
rs504963	*FUT2*	0.472	0.001	0.967	0.965

[a]Kendall's correlation coefficient.
[b]Percentile rank relative to the distribution of all SNPs.
[c]Percentile rank relative to the distribution of allele frequency-matched SNPs.

The strongest correlation between BGA SNP allele frequency and pathogen richness was obtained for rs900971 in *SLC14A1* (Fig. 1; similar representations for all SNPs in Table 1 are available as Supplemental Fig. 1). In order to verify that environmental variables correlating with pathogen richness (Guernier et al. 2004) did not determine the association signal with BGA genes, we calculated the mean temperature and maximum precipitation rate for geographic locations corresponding to HGDP-CEPH populations; none of the SNPs reported in Table 1 significantly correlated with either variable (data not shown).

The identification of correlations between specific environmental variables and allele frequencies has been regarded as a strategy complementary to common population genetic approaches for the detection of selection signatures (Hancock et al. 2008). All such analyses rely on the assumption that the environmental variable we measure nowadays has changed little over human history and that gene flow due to recent admixture has had a minor impact on human genetic diversity. In this case, we implicitly assumed that the number of different pathogen species per country has been maintained proportionally unchanged along human evolutionary history. Although an oversimplification, this might not be so different from the reality, given that climatic variables have been shown to be of primary importance in driving the distribution of human pathogens (Guernier et al. 2004). As for gene flow, the influence of recent admixture in most populations is considered to be modest (Li et al. 2008), as also demonstrated by the good relationship between population genetic diversity and distance from Africa (Handley et al. 2007).

Our data therefore indicate that the allele frequencies of a subset of BGA genes vary with pathogen richness, supporting the vision whereby these loci affect the susceptibility to infectious diseases. This hypothesis had previously been formulated for *ABO* and *FUT2* (Greenwell 1997; Hill 2006; Casanova and Abel 2007), while in the case of *GYPC, DARC*, and *SLC4A1* the ability of specific alleles to modulate infection susceptibility has been demonstrated for malaria (Moulds and Moulds 2000). Also, in the case of *AQP3*, modulation of malaria severity can be hypothesized since *AQP3* represents the major channel for glycerol transport in

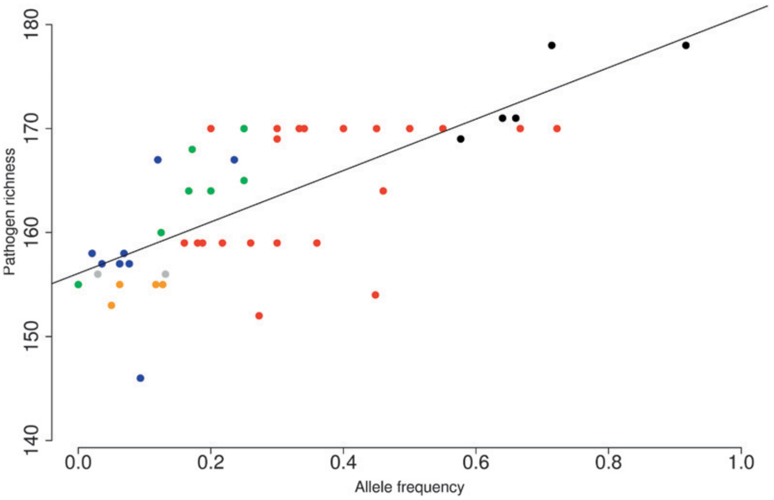

Figure 1. Correlation between pathogen richness and allele frequency for rs900971 in *SLC14A1*. Populations from different broad geographic areas are coded by different colors: (green) Sub-Saharan Africa, (black) America, (red) Asia, (blue) Europe, (orange) Middle East, and (gray) Oceania.

human erythrocytes (Roudier et al. 2002), and mice knockout for *Aqp9*, a related glycerol transporting aquaporin, display increased survival to *P. berghei* (Liu et al. 2007). One possibility to explain the observed associations is that pathogen richness has co-varied with malaria prevalence and that nucleotide variations, even different from those previously reported to confer resistance in *SLC4A1* and *GYPC*, affect *Plasmodium* entry, spread, or rosetting. Conversely, no association with infectious disease predisposition has ever been reported for *SLC14A1, ERMAP, C1GALT1*, and *GCNT2*; yet these genes encode either glycotransferases or surface glycosylated proteins, suggesting that carbohydrate determinants might affect pathogen attachment and entry.

Another possibility involving *SLC14A1* variants is that the association with pathogen richness reflects some important aspect of urea metabolism during infection; indeed the gene codes for an urea transporter and the intracellular availability of urea has been shown to be a limiting factor for the ability of *Mycobacterim bovis* to attenuate expression of MHC class II molecules during macrophage infection through urease-induced alkalinization of intracellular compartments (Sendide et al. 2004).

As for *CD44*, it has been shown to act as a receptor for group A *Streptococcus* (Cywes et al. 2000), *Mycobacterium tuberculosis* (Leemans et al. 2003), and *Escherichia coli* in urinary tract infections (Rouschop et al. 2006).

Population genetics analysis of BGA genes

Given the results obtained above and the premise whereby BGA genes might represent selective targets, we wished to verify whether selection signatures could be identified at BGA genes. To this aim we exploited the fact that 22 out of 38 loci involved in BGA specification have been included in the SeattleSNPs program so that resequencing data (although with some gaps) in at least two populations are available; in particular, all data refer to one population with European ancestry (EA) and one with African ancestry (either Yorubans [YRI] or African American [AA]). From the SeattleSNP gene list we excluded *ABO*, which has been previously studied (Saitou and Yamamoto 1997; Calafell et al. 2008), *A4GALT*, *XK*, and *ART4* due to poor resequencing coverage, and *KEL*, because of its being located in a region subjected to a selective sweep possibly driven by the nearby *TRPV6* locus (Akey et al. 2006). The following genes were left for analysis: *AQP1*, *AQP3*, *ACHE*, *BSG*, *B3GALNT1* (previously *B3GALT3*), *CD55* (previously *DAF*), *CD151*, *SLC4A1*, *ICAM4*, *FUT3*, *FUT2*, *FUT1*, *BCAM*, *ERMAP*, *GYPC*, *SEMA7A*, and *SLC14A1*.

With the aim of identifying loci that have been subjected to natural selection, and following the conundrum whereby selection signatures might extend over relatively short gene regions (due to the action of mutation and recombination; Wiuf et al. 2004; Bubb et al. 2006), we applied a sliding window approach to all BGA genes (except for *ACHE*, due to its small size and *FUT2*, as detailed below) and calculated population genetic differentiation, measured as $F_{ST.}$ Under the assumption of neutrality, F_{ST} is determined by demographic history (i.e., genetic drift and gene flow), which affects all loci similarly. We therefore calculated the 2.5th and 97.5th percentiles in the distribution of F_{ST}-values obtained for sliding windows across SeattleSNPs genes (see Methods for details) and searched for BGA gene regions that display unusually high or low population differentiation. Overall, 8.3% of sliding windows deriving from the 17 BGA genes displayed exceedingly high or low F_{ST}-values; estimation of an empirical probability (see Methods) to obtain an equal or higher fraction of outliers in windows deriving from SeattleSNP genes yielded a *P*-value of 0.19. These data indicate that an excess of unusual F_{ST}-values can be observed for BGA genes, with the failure to reach statistical significance being likely due to the presence of other non-neutrally evolving genes in the SeattleSNP data set (which mainly gathers genes involved in inflammatory processes).

BGA gene regions displaying unusual F_{ST}-values were further studied by application of population genetics statistics. In particular, widely used test include Tajima's *D* (Tajima 1989) and Fu and Li's *D** and *F** (Fu and Li 1993). Tajima's *D* (D_T) tests the departure from neutrality by comparing two nucleotide diversity indexes: θ_W (Watterson 1975), an estimate of the expected per site heterozygosity, and π (Nei and Li 1979), the average number of pairwise sequence nucleotide differences. Positive values of D_T indicate an excess of intermediate frequency variants and are a hallmark of balancing selection; negative D_T-values indicate either purifying selection or a high representation of rare variants as a result of a selective sweep. Fu and Li's *F** and *D** are also based on SNP frequency spectra and differ from D_T in that they also take into account whether mutations occur in external or internal branches of a genealogy. Since population history, in addition to selective processes, is known to affect frequency spectra and all related statistics; we performed coalescent simulations using a calibrated population genetics model that incorporates demographic scenarios (Schaffner et al. 2005). Also, in order to disentangle the effects of selection and population history, we exploited the conundrum whereby selection acts on a single locus while demography affects the whole genome: As a control data set we therefore calculated diversity parameters and test statistics for 5 kb windows deriving from 238 genes resequenced by the NIEHS program (see Methods for details). A similar comparison with

SeattleSNPs gene data is provided in Supplementary File 2 (Supplemental Table 2). Sliding window analyses identified gene regions showing unusual F_{ST}-values (Supplemental Fig. 3), which were selected for further study, as reported in the following paragraphs. In addition, for the remaining BGA loci we calculated summary statistics (D_T, D^*, and F^*) for the entire gene region and unusual values were found for *BSG* (detailed below) and *FUT1*. The latter was not further analyzed as high homology with other gene family members and a pseudogene suggested that gene conversion events might affect the result (this was not the case for *FUT2* since we focused on the promoter region, as detailed below).

CD55 (Cromer system)

A sliding-window analysis along the *CD55* gene (OMIM no. +125240; referred to as *DAF* in the SeattleSNPs database) revealed the presence of a region encompassing nucleotides ~9000–19,000 showing exceedingly low F_{ST}-values (Supplemental Fig. 3). Both nucleotide diversity estimates and test statistics (Table 2) revealed no significant departure from neutrality for both, AA and EA, yet D_T and Fu and Li's D^* and F^* ranked relatively high in the distribution of 5 kb windows from NIEHS genes (see Supplemental Table 2 for a comparison with SeattleSNPs genes).

Under neutral evolution, the amount of within-species diversity is predicted to correlate with levels of between-species divergence, since both depend on the neutral mutation rate (Kimura 1983). The HKA test (Hudson et al. 1987) is commonly used to verify whether this expectation is verified. Here we performed a maximum likelihood HKA test (MLHKA) by comparing the *CD55* region to 16 neutrally evolving genes (see Methods for details): A significant result was obtained for both AA and EA (Table 3).

Therefore, we wished to study the genealogy of *CD55* haplotypes in the region and to this aim a neighbor-joining network was constructed. Two major clades separated by long branch lengths are evident (Fig. 2), each containing common haplotypes. In order to estimate the TMRCA (time to the most recent common ancestor) of the two haplotype clades, we applied a phylogeny-based method (Bandelt et al. 1999) based on the measure ρ, the average pairwise difference between the two haplotype clusters. ρ resulted in a value equal to 13.28, so that, using a mutation rate based on 50 fixed differences between chimpanzee and humans, and a separation time of 6 million years (Myr) (Glazko and Nei 2003), we estimated a TMRCA of 3.19 Myr (SD = 673 Kyr). Given the low recombination rate in the region, we wished to verify this result using GENETREE, which is based on a maximum-likelihood coalescent analysis (Griffiths and Tavare 1994, 1995). The method assumes an infinite-site model without recombination and, therefore, haplotypes and sites that violate these assumptions need to be removed: in this case, only three single segregating sites had to be removed. The resulting gene tree, rooted using the chimpanzee sequence, is partitioned into two deep branches (Supplemental Fig. 4). A maximum-likelihood estimate of θ (θ_{ML}) of 9.2 was obtained, resulting in an estimated effective population size (N_e) of 22,000, a value comparable to most figures reported in the literature (Tishkoff and Verrelli 2003). Using this method, the TMRCA of the *CD55* haplotype lineages amounted to 2.61 Myr (SD = 552 Kyr). Such deep coalescent time is unusual, as estimates for neutrally evolving autosomal loci range between 0.8 Myr and 1.5 Myr (Tishkoff and Verrelli 2003).

Overall these data strongly support the idea that the *CD55* region we analyzed has evolved under long-standing balancing selection. This gene portion covers roughly 10 kb surrounding exon 6–7 and contains four DNase I hypersensitive sites in $CD4^+$ T cells (Boyle et al. 2008); five intermediate frequency SNPs (rs6700079, rs2184476, rs1507760, rs10746462, and rs10746463) located along the branch separating the two haplogroups lie within DNase

Table 2. Summary statistics for selected BGA regions

							D			D*			F*		
Gene	L[a]	P[b]	N[c]	S[d]	θ[e]	π[f]		P[g]	Rank[h]		P[g]	Rank[h]		P[g]	Rank[h]
CD55	9.5	AA	48	48	11.39	12.36	0.30	0.11	0.88	−0.52	0.46	0.45	−0.27	0.29	0.56
		EA	46	34	8.14	10.63	1.04	0.14	0.82	0.68	0.16	0.77	0.96	0.11	0.81
CD151	1.6	YRI	48	15	21.12	30.70	1.40	0.022	0.99	0.66	0.15	0.88	1.08	0.048	0.96
		AA	46	14	19.91	35.94	2.48	0.0014	>0.99	0.59	0.18	0.87	1.44	0.015	>0.99
		EA	46	13	18.49	26.01	1.24	0.11	0.85	0.51	0.31	0.71	0.89	0.19	0.78
		AS	40	12	17.63	10.74	−1.20	0.095	0.15	−3.11	0.0067	0.029	−2.94	0.013	0.043
FUT2 pm[i]	5.8	YRI	48	40	15.58	30.06	3.18	<0.0001	>0.99	1.29	0.0054	0.98	2.34	<0.0001	>0.99
		EA	46	8	3.15	1.63	−1.38	0.046	0.13	−0.80	0.27	0.27	−1.14	0.17	0.24
FUT2 cds[j]	3	YRI	48	26	19.52	29.94	1.76	0.0033	>0.99	−0.60	0.55	0.42	0.27	0.18	0.77
		EA	46	20	15.16	24.88	2.07	0.018	0.97	0.60	0.23	0.73	1.30	0.068	0.92
SLC14A1	9.9	YRI	48	86	19.57	18.52	−0.19	0.26	0.64	0.61	0.045	0.86	0.37	0.073	0.80
		AA	48	87	19.80	22.89	0.55	0.051	0.94	1.58	<0.0001	>0.99	1.42	0.0005	0.99
		EA	46	62	14.25	21.97	1.91	0.015	0.96	0.89	0.087	0.83	1.50	0.022	0.95
		AS	40	57	13.54	24.19	2.82	0.0007	>0.99	1.51	0.0074	>0.99	2.34	0.0006	>0.99
BSG	11.3	YRI	48	80	15.95	16.45	0.11	0.24	0.83	0.38	0.24	0.84	0.34	0.22	0.84
		EA	46	81	16.31	11.91	−0.96	0.18	0.18	−2.08	0.052	0.071	−2.00	0.054	0.076

[a] Length of analyzed resequenced region (kb).
[b] Population.
[c] Sample size.
[d] Number of segregating sites.
[e] θ_W estimation per site ($\times 10^{-4}$).
[f] π estimation per site ($\times 10^{-4}$).
[g] *P*-values obtained by applying a calibrated population genetics model, as described in the text.
[h] Percentile rank relative to the distribution of 5 kb deriving from 238 NIEHS genes.
[i] Promoter region.
[j] Coding region.

Table 3. MLHKA test results for BGA regions

		MLHKA	
Gene	Population	*k*	*P*-value
CD55	AA	3.56	0.0013
	EA	3.25	0.0035
CD151	YRI	3.65	0.0057
	AA	3.08	0.015
	EA	4.46	0.0031
	AS	4.30	0.0042
FUT2 pm	YRI	1.86	0.098
	EA	0.43	0.017
SLC14A1	YRI	2.68	0.0044
	AA	2.79	0.0018
	EA	2.51	0.0066
	AS	2.84	0.0059

k, Selection parameter.

I hypersensitive sites. Since DNase I hyperaccessibilty is thought to be a hallmark of active *cis*-regulatory regions (Gross and Garrard 1988; Felsenfeld and Groudine 2003), these variants might represent good candidates as functional SNPs with a role in transcriptional regulation of *CD55*. Importantly, another variant (rs6700168) located in this genomic portion was found to correlate with pathogen richness (Table 1) and it lies along the branch separating the two haplotype clusters (Fig. 2). In order to verify whether heterozygote advantage might underlie the action of balancing selection we calculated the observed over expected heterozygosity for rs6700168 and verified whether this ratio varied with pathogen richness. Since this was not the case, we suggest that the maintenance of the two haplotype lineages is not due to overdominance but possibly to antagonistic selection (see below).

CD55 (also known as *DAF*, decay-accelerating factor) is a complement-regulatory protein expressed by most cell types, which protects host tissues from damage by the autologous complement system (Nicholson-Weller and Wang 1994). Previous studies have indicated that the membrane-anchored form of *CD55* serves as a receptor for very common human pathogens, such as Dr$^+$ *E. coli* (Nowicki et al. 1993), coxsackieviruses B1, B3, and B5 (Shafren et al. 1995), and echovirus 7 (Clarkson et al. 1995), suggesting that decreased or abolished *DAF* expression might confer decreased susceptibility to these infectious agents. Total absence of *CD55* (Inab phenotype) is very rare in humans (Blumenfeld and Patnaik 2004) and associates with no overt phenotype. Yet, other observations point to a possible role of the gene in fertility and pregnancy: *CD55* is dynamically regulated during the menstrual cycle (Young et al. 2002) and it is highly expressed at the feto–maternal interface (Sood et al. 2006); moreover, reduced *DAF* expression has been associated with luteal phase defect of the endometrium associated with infertility or pregnancy loss. Also, mice lacking *DAF* are more susceptible to autoimmune manifestations (Kaul et al. 1995).

These evidences might therefore suggest that regulation of *CD55* expression levels, either in a cell-type- or stage-dependent fashion might affect vital processes, such as reproduction and immunity. Also, the lack of evidence supporting heterozygote advantage and the phenotype of *cd55*$^{-/-}$ mice possibly suggest that balancing selection ensues from antagonist selection trading-off resistance to infection with autoimmune phenomena. Obviously, other hypotheses are possible (e.g., adaptation to variable environmental conditions with special reference to different environmental pathogens) and further studies on the biological function of *CD55* will be instrumental in addressing this issue.

CD151 (RAPH system)

A sliding window analysis along *CD151* (OMIM no. *602243) indicated the 3′ gene region displays reduced population differentiation and exceedingly low F_{ST}-values are observed in a region roughly corresponding to the terminal region extending from exon 6 to the 3′ UTR (Supplemental Fig. 3). Nucleotide diversity (both θ_W and π), in this restricted region, ranked above the 97.5th percentile in the distribution of 5 kb windows deriving from NIEHS genes (Supplemental Table 1) for both EA and YRI. Summary statistics revealed significantly positive values for D_T, D*, and F* in YRI, but not in EA. In order to further investigate the possible departure from neutrality in other human populations, the same region was resequenced in two additional samples: Asians (AS) and AA. As shown in Table 2, significantly positive test statistics were obtained for populations of African ancestry but not for Asians and Europeans. In the case of AS, negative values of summary statistics are due to the presence of a single highly divergent haplotype (Fig. 3).

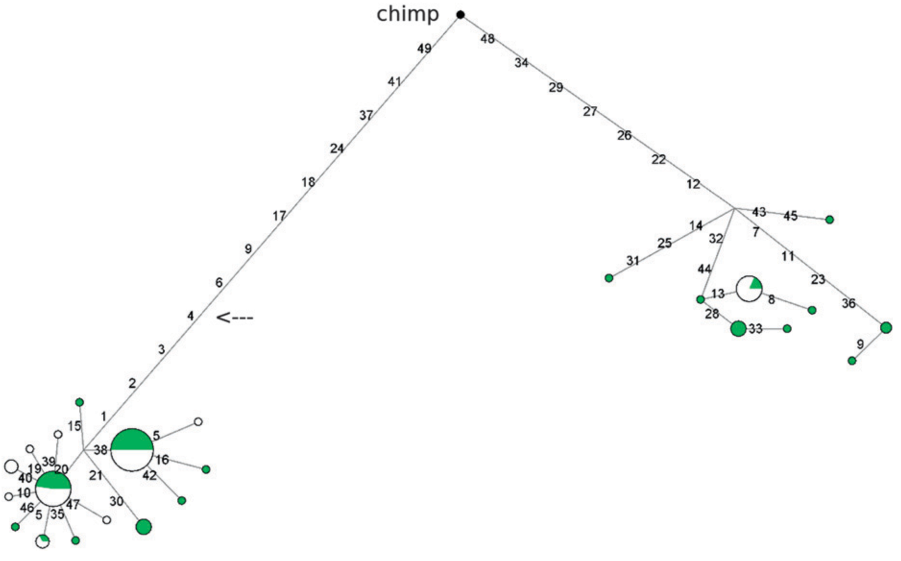

Figure 2. Genealogy of *CD55* haplotypes reconstructed through a median-joining network. The analysis corresponds to the gene region spanning nucleotides ~9500–18,300 (as described in the text). Each node represents a different haplotype, with the size of the circle proportional to the haplotype frequency. Nucleotide differences between haplotypes are indicated on the branches of the network. Circles are color-coded according to population (green, AA; white, EA). The chimpanzee sequence is also shown. The arrow shows the position of rs6700168 (Table 1). Note that the relative position of mutations along a branch is arbitrary.

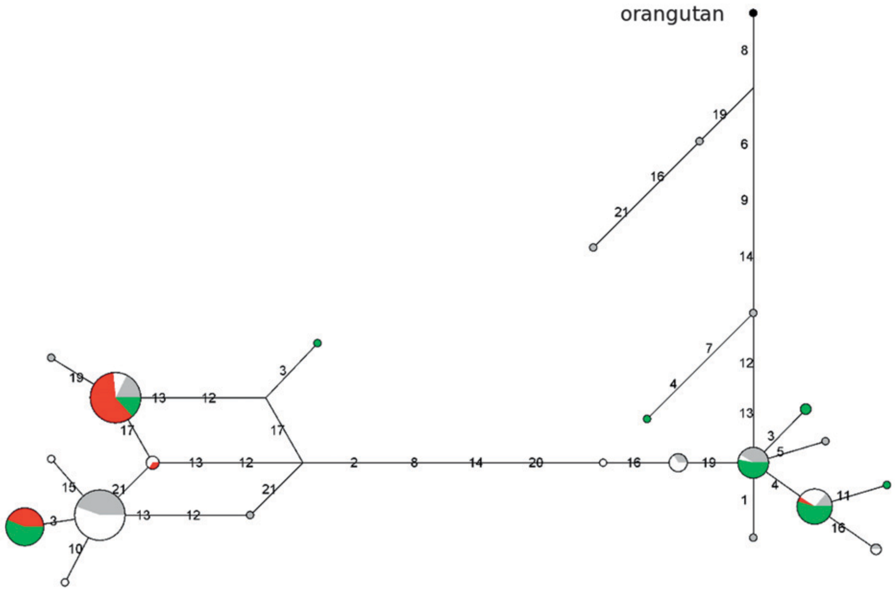

Figure 3. Genealogy of *CD151* haplotypes reconstructed through a median-joining network. The analysis corresponds to the gene region spanning nucleotides 9100–10,400. Population color codes are as follows: (green) AA; (white) EA; (red) AS; (gray) YRI.

Application of the MLHKA test, using 16 neutrally evolving genes (see Methods), rejected the hypothesis of neutrality for all populations (Table 3).

Next, we wished to examine haplotype genealogy for the terminal *CD151* gene region and, to this aim, a median-joining network was constructed (Fig. 3). The topology of this network was relatively unambiguous showing two major clades, each containing common haplotypes, separated by long branch lengths. Calculation of the TMRCA ($\rho = 8.55$; fixed differences = 58 using *Pongo pygmaeus*) yielded an estimate of 3.83 Myr (SD = 1.06 Myr), again a deep coalescent time compared to neutral loci. As reported above for *CD55*, this result was verified using GENETREE and an estimated TMRCA of 2.14 Myr was obtained (SD = 643 Kyr, $\theta_{ML} = 1.9$, $N_e = 13{,}722$; Supplemental Fig. 5). In analogy to *CD55*, these features suggest the action of long-standing balancing selection in African populations.

The analyzed *CD151* region covers the last four coding exons and the 3′ UTR; most variants are located in noncoding regions, with the majority of intermediate frequency SNPs falling within the UTR. Analysis of known functional elements in the 3′ UTR was performed using UTRscan and no SNPs were found to affect predicted motifs. Conversely, a search for microRNA target sites (miRBase) indicated that one variant, namely rs1130698, falls within the highest scoring predicted target site. In particular the T allele changes a G–C pairing between the *CD151* UTR sequence and hsa-miR-940 to a G–U wobble; unfortunately, little is known about the expression pattern of hsa-miR-940, except for the fact that it was cloned from cervical cell lines (Lui et al. 2007).

CD151, a member of the tetraspanin protein family involved in cell adhesion and motility, is expressed in most human tissues (Fitter et al. 1995). Mutations of *CD151* in humans result in nephropathy with epidermolysis bullosa and deafness (Karamatic Crew et al. 2004), while different phenotypes have been reported for $cd151^{-/-}$ mice, including abnormal hemostasis (Wright et al. 2004), defective wound healing (Cowin et al. 2006), and renal defects (Sachs et al. 2006). Members of the tetraspanin family have been implicated in virus infection in animals and humans; in particular, different tetraspanins have been shown to act as receptors for HCV (Pileri et al. 1998), HIV (von Lindern et al. 2003), canine distemper virus (Loffler et al. 1997), feline leukemia virus (Willett et al. 1994), and porcine reproductive and respiratory syndrome virus (Shanmukhappa et al. 2007); yet a recent report has also shown that members of the tetraspanin family, including *CD151*, protect human macrophages from HIV-1 and vescicular stomatitis virus infection, possibly by blocking virion binding/uptake (Ho et al. 2006).

Whether the maintenance of balancing selection at the *CD151* locus is pathogen-driven remains to be elucidated, and unfortunately no SNP mapping to the gene has been typed in the HGDP-CEPH panel; it is worth noting that besides its possible direct role in predisposing to infections (by acting as a viral receptor/binding factor), its function in wound healing (Cowin et al. 2006) might also be regarded as linked to pathogens and their prevalence, in that the risk of wound infection likely depends on how long the healing process takes to completion.

FUT2 (Lewis system)

In humans two alpha (1,2)-fucosyltransferases, encoded by the paralogous *FUT1* and *FUT2* genes, determine expression of the human H antigen, a precursor of blood group A and B antigens.

The two genes differ in substrate specificities and tissue expression (Costache et al. 1997): *FUT1* (H enzyme, H/h system) is responsible for the expression of H antigen in red cells and vascular endothelia, whereas the *Se* enzyme (encoded by *FUT2*, Lewis system, OMIM no. +182100) is responsible for the synthesis of the same antigen in secretory glands and the intestinal mucosa; individuals referred to as "secretors" (*Se*) have at least one functional *FUT2* allele.

Common *FUT2* null alleles are present in many populations; in particular a frequent null allele (se^{428}) is responsible for most nonsecretor phenotypes in Europe and Africa, while a missense mutation (se^{385}) is widespread in East Asians (Kelly et al. 1995; Koda et al. 1996; Liu et al. 1998). Interestingly, the coding region of *FUT2* has previously been hypothesized to be subjected to balancing selection, possibly under an overdominance regime (Koda et al. 2001).

In the case of *FUT2* we did not perform a sliding window analysis as described for the above genes due to extensive resequencing gaps. Rather, we divided the gene in three major portions: coding exon, intron, and 5′ upstream region (10 kb upstream the transcription start site, thereafter referred to as putative promoter). In line with previous findings, the coding exon displayed high nucleotide diversity and positive statistics (Table 2), while we verified that low levels of nucleotide variation characterize the only intron (not shown). Interestingly, an unusual pattern was observed at the putative promoter region: as shown in Table 2, YRI displayed high values of $\theta_{W,}$ while EA presented low nucleotide diversity (percentile rank of $\theta_W = 0.18$). Calculation of F_{ST} yielded a high value of 0.45, corresponding to a percentile rank of 0.977 in the distribution of SeattleSNPs gene windows and being 20-fold greater than population differentiation calculated for the coding region ($F_{ST} = 0.022$).

Summary statistics for the putative promoter revealed deviation from neutrality in YRI, since all tests yielded significantly positive values (Table 2); conversely, statistics for EA resulted in negative values, although significance was only obtained for D_T. Human/chimpanzee divergence in this gene region amounted to 1.35%, a value higher than the genome average (average = 1.06%, SD = 0.25%; Chimpanzee Sequencing and Analysis Consortium 2005) and greater than that of control loci used in the MLHKA test; the latter gave no significant results for YRI, while a reduction of polymorphism compared to intraspecific divergence was evidenced for EA (Table 3). The greater than average divergence and high polymorphism level observed for YRI might be consistent with the region having low sequence constraints, resulting in an increase of both divergence and diversity; yet, this hypothesis does not fit the EA data whereby low diversity is observed; moreover, the high population differentiation we observed can hardly be reconciled with a neutral pattern of evolution.

Low diversity values and negative statistics are consistent with both purifying and directional selection; Fay and Wu's *H* (Fay and Wu 2000) is usually applied to distinguish between these possibilities, since significantly negative *H*-values indicate an excess of high-frequency derived alleles, consistent with directional selection. *H* equaled −4.66 in EA with a borderline *P*-value of 0.062 (calculated using the calibrated model). It should be noted that the interpretation of *H* can be complicated by the fact that the power of this statistic to detect selection is poor when the sweep is relatively old (Przeworski 2002) and population structure can result in significantly negative *H* statistics (Przeworski 2002).

Reconstruction of haplotype genealogy for the *FUT2* putative promoter using yielded a topology with two major clades separated by long branch lengths (Fig. 4); consistent with the high degree of geographic structure, all European haplotypes cluster with the same haplogroup while African chromosomes are divided in the two clades. Calculation of the TMRCA (ρ = 13.10; fixed differences = 79, using chimpanzee) yielded an estimate of 1.99 Myr (SD = 410 Kyr). A similar TMRCA was estimated with the use of GENETREE (TMRCA = 1.70 Myr, SD = 375 Kyr, $\theta_{ML} = 3.4$, $N_e = 8681$; Supplemental Fig. 6). Construction of a haplotype genealogy for the coding region (data not shown) resulted in a TMRCA of 3 Myr, in agreement with previous findings (Koda et al. 2001).

Overall, the data presented above are consistent with the presence of a selected variant/haplotype in the promoter region of *FUT2*; this is in line with a recent report indicating that distinct promoter haplotypes have an effect on the gene transcription levels (Soejima and Koda 2008). In the case of EA, the statistics we performed did not allow a firm rejection of the neutral model; in part this might be due to the small number of SNPs in the region (only eight) which reduces the power of all tests; also, failure to reject neutrality might be accounted for by the pattern being a relic of older selective events.

In the case of YRI, we consider that our observations might be consistent with the presence of a balanced polymorphism. This raises the possibility that the signatures we obtained at the promoter region are due to hitchhiking and linkage disequilibrium (LD) with the coding exon. Nonetheless, different observations suggest that this is not the case. First, calculation of *D′* between the se^{428} variant and common SNPs in the putative promoter revealed a maximum value of 0.27, indicating low LD, in agreement with a previous report (Soejima and Koda 2008). Second, summary statistics yielded stronger results for the putative promoter compared to the coding exon. Third, although hitchhiking has the potential to affect large genomic regions, the signatures of balancing selection are predicted to extend over relatively short distances (Wiuf et al. 2004; Bubb et al. 2006); as an example, the high nucleotide diversity that characterizes the second exon of MHC loci decays rapidly in flanking intronic sequences (Cereb et al. 1997; Fu et al. 2003) and neighboring exons (Takahata and Sata 1998). This suggests that the departure from neutrality and the high level of nucleotide diversity we observe in the *FUT2* putative promoter region is not merely a result of hitchhiking with the coding exon, given the 7 kb separating the transcription start site from the second exon.

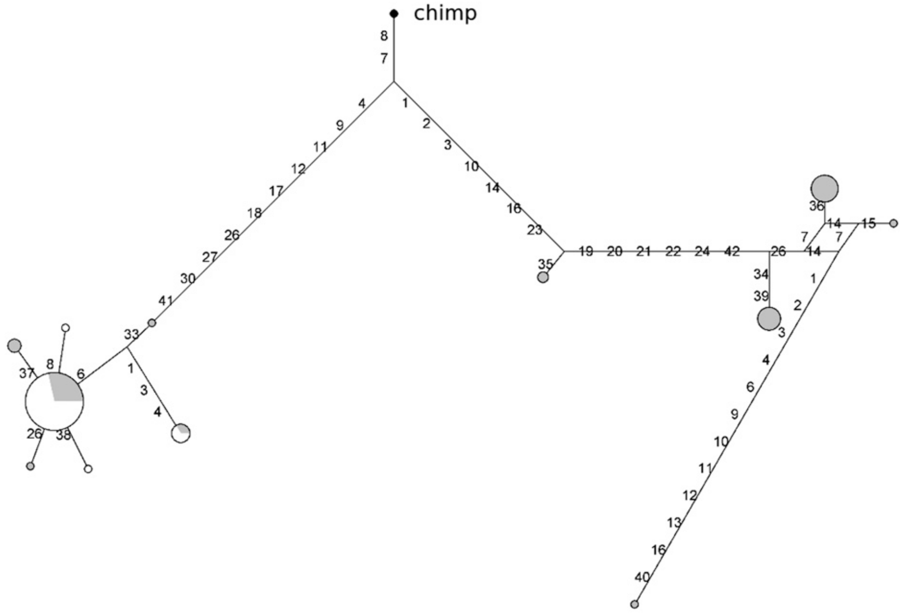

Figure 4. Median-joining network for the putative promoter region of *FUT2*. The analysis corresponds to the gene region spanning nucleotides 7400–17,800. Population color codes are as follows: (white) EA; (gray) YRI.

As reported above, a recent study (Soejima and Koda 2008) of the *FUT2* proximal promoter region indicated that nucleotide diversity patterns differ between African and non-African populations, and the authors identified two common haplotypes with different cell-type specific activities. These observations raise the interesting possibility that balancing selection at the *FUT2* promoter region might result from overdominance due to differential activity of the two promoter haplotypes in different tissues. The "secretor" status has been associated with increased susceptibility to infection by caliciviruses (Lindesmith et al. 2003), HIV (Ali et al. 2000), and respiratory viruses (Raza et al. 1991); yet secretor subjects also display advantages compared to nonsecretors, such as lower susceptibility to urinary tract and *Candida* infections, and increased protection against *Neisseria meningitis* and *Streptococcus* (Haverkorn and Goslings 1969; Blackwell et al. 1990). Also, situations exist where the secretor status might underlie a double-faceted situation. One example involves *Campylobacter jejuni* infection (the most common cause of bacterial diarrhea): the pathogen exploits H antigens for tethering to the intestinal mucosa, but at the same time alpha (1,2)-linked fucosyloligosaccharides in human milk inhibit *Campylobacter* infection by competing with intestinal cell surface receptors (Ruiz-Palacios et al. 2003). As a result, a breast-fed infant is expected to be at variable risk of infectious diarrhea depending on his/her intestinal expression of H antigens and his/her mother secretion of the same molecule in milk; in this scenario, maximization of *FUT2* expression in lactating epithelia might be extremely important in providing immunization to newborns. Indeed, different oligosaccharide species in human milk form part of the innate immune system with activity against different pathogens (Newburg et al. 2005), and fucosyloligosaccharides containing alpha (1,2)-linked fucose are prevalent (Chaturvedi et al. 2001). Women who are nonsecretors do not express measurable 2-linked fucosyloligosaccharides and the amount of milk fucosyloligosaccharides varies even among secretors (Chaturvedi et al. 2001), possibly suggesting the presence of genotype differences responsible for such variation (Chaturvedi et al. 2001). Since diarrhea represents a very common cause of mortality in newborns throughout the world, the adaptive significance of decreasing the chance of infection in breast-fed infants is evident. Therefore, maintenance of the advantages conferred by the secretor status, while modulating the levels of glycotransferase activity in a cell-type-dependent fashion, might represent a beneficial strategy in specific circumstances. Obviously, other explanations for the maintenance of different *FUT2* promoter haplotypes are possible, and further studies will be required to analyze the activity of *FUT2* promoter haplotypes. Unfortunately, no SNP located in the putative promoter region of *FUT2* was available to test association with pathogen richness or verify whether a heterozygote excess could be observed in specific geographic locations. Conversely, significantly associated SNPs are located in the coding exon (rs602662 and rs485186) or 3′ UTR (rs504963) and are in full LD with the null se^{428} allele in both EA and AA (in all cases, $D' = 1$, $P < 0.001$). No correlation was observed between the observed/expected heterozygosity ratio for these SNPs and pathogen richness, suggesting that, although pathogens have exerted a selective pressure on the gene and balancing selection has been operating, the underlying explanation is not accounted for by overdominance. This might be expected if the null allele is thought of as the selected variant: Heterozygotes are secretors and they are not expected to have an advantage compared to subjects carrying two active alleles. One possibility is that secretors and nonsecretors experience advantages or disadvantages depending on variable environmental conditions in terms of pathogen prevalence, since they display different susceptibility to diverse pathogen types. Alternatively, as suggested above, more complex scenarios can be envisaged that also take into account promoter variants.

SLC14A1 (Kidd system)

Sliding window analysis of *SLC14A1* (OMIM no. *111000) revealed an extended region of about 6.3 kb showing high levels of population differentiation (Supplemental Fig. 3). The single variant (Asp280Asn) responsible for the common *JK*A/JK*B* antigens (Blumenfeld and Patnaik 2004) is located ~6 kb downstream and, with the aim of analyzing the evolutionary history of the gene, we decided to resequence the entire region in YRI and AS with the exception of a small central gap of 2 kb (Supplemental Fig. 3). Two novel nonsynonymous variants were identified, Val10Met and Val76Ile, and both were present in the same three AA subjects.

Summary statistics and diversity parameters for the four populations (Table 2) revealed high levels of polymorphism and allowed rejection of neutrality for AA, AS, and EA, while borderline values were obtained for YRI.

Application of the MLHKA test, as described above, rejected the hypothesis of neutrality for all populations (Table 3). Construction of the median-joining network is recommended when regions displaying low recombination are being analyzed; in the

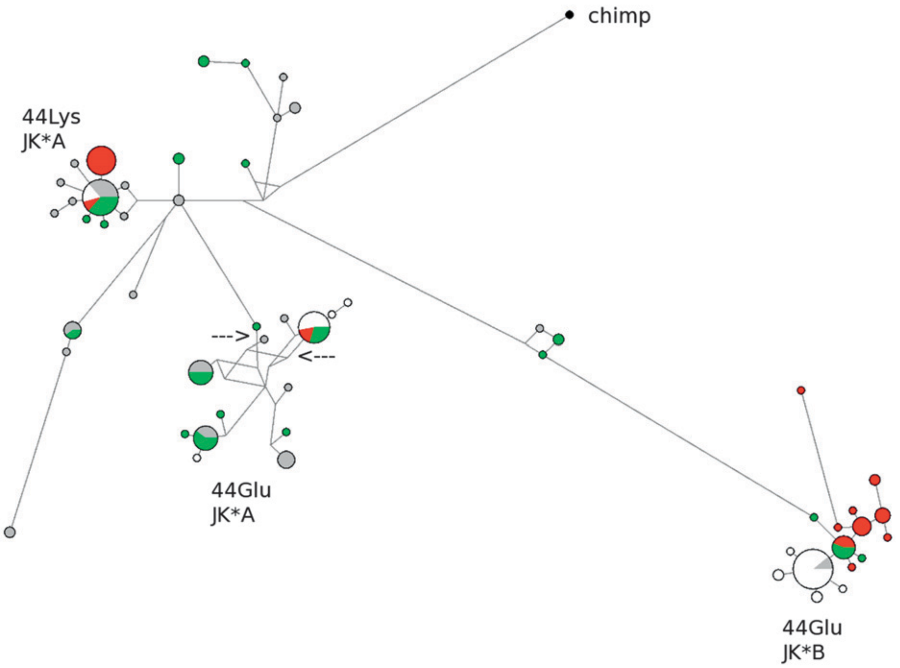

Figure 5. Genealogy of *SLC14A1* haplotypes reconstructed through a median-joining network. The analysis corresponds to the gene region spanning nucleotides 4887–17,350. Population color codes are as follows: (green) AA; (white) EA; (red) AS; (gray) YRI. The allelic status at amino acid position 44 and 280 (*JK*A/JK*B*) is reported for the three major clusters. The two arrows denote the position of rs10853535 and rs692899, which correlate with pathogen richness (Table 1).

case of *SLC14A1*, the gene region carrying the Asp280Asn polymorphism displays low LD with the more 5′ region (Supplemental Fig. 7); yet, we decided to calculate the network over the entire region so that the relative distribution of chromosomes carrying the *JK*A/JK*B* variants could be visualized (Fig. 5); conversely, TMRCA estimate was performed using GENETREE and for this analysis only variants in linkage disequilibrium were included (Supplemental Fig. 7). In both cases, three major haplotype clades are evident and TMRCA estimated equaled 2.28 Myr (SD = 283 Kyr, θ_{ML} = 12, N_e = 28,800). The median joining network shows that a long branch separates haplotypes carrying the *JK*B* allele from *JK*A*, while a nonsynonymous Glu44Lys SNP might be regarded as the selected variant maintaining the two closer clusters carrying *JK*A* (Fig. 5). It is interesting to notice that two variants located in this gene region (rs10853535 and rs692899) correlate with pathogen richness (Table 1); one of them lies on the branch leading to the haplotype cluster carrying the *JK*A* and 44Glu alleles, while the second is internal to this same cluster and defines a smaller haplotype group (Fig. 5).

Interestingly, for both these variants the observed over expected heterozygosity ratio significantly correlated with pathogen richness (rs692899, τ = 0.327, P = 0.0012; rs10853535, τ = 0.311, P = 0.0019, Supplemental Fig. 2), possibly suggesting that the two subclades carrying the *JK*A* allele are maintained by overdominance. Conversely, we found no heterozygote excess for rs900971, which showed the strongest correlation with pathogens among all BGA SNPs (Table 1). This variant is located further downstream the Asp280Asn SNP and displays low linkage disequilibrium with the balancing selection region. It is therefore tempting to speculate that different variants in *SLC14A1* have been subjected to pathogen-driven selection under different regimes that might include heterozygote advantage and, possibly, directional selection.

Altogether, the data reported above concur with the idea that multiallelic balancing selection has shaped the evolutionary history of *SLC14A1,* although several issues remain to be clarified. In particular, the possible role of urea metabolism in relation to pathogen resistance has been briefly mentioned above as a possible explanation for selection at this locus, but current knowledge on this issue is too limited to warrant extensive speculation. Moreover, consistent with the biological function of *SLC14A1*, Kidd-null subjects and knockout mice display mild urinary concentrating defects and greater urine output (Sands et al. 1992; Yang et al. 2002). This observation raises the possibility that, together with pathogen-driven selection, the transporter might also have adapted to climatic variables, possibly driven, for example, by the necessity to spare water in hot dry climates. In fact, we did not find any SNP in *SLC14A1* to correlate with climatic variables, such as mean temperature and maximum precipitation rate. Yet, the effect might be confounded by pathogen-driven selection or the power to detect a correlation might vary depending on the environmental variable, as previously suggested (Hancock et al. 2008).

BSG (OK system)

Calculation of nucleotide diversity parameters and summary statistics for the whole *BSG* gene (OMIM no. *109480) revealed an unusual pattern in EA. In both this population and in YRI we observed a θ_W of 16×10^{-4}, a value higher than the 97.5th percentile in EA (Supplemental Table 1). Yet, while in YRI relatively high values for Fu and Li's D^* and F^* were obtained, all statistics were negative in EA with borderline significance (Table 2). Closer examination indicated that the negative statistics in Europeans are due to the presence of a single highly divergent haplotype carrying 24 singletons. We therefore verified whether this haplotype was present in the African sample and identified six additional chromosomes carrying closely related haplotypes. We next constructed a median-joining network of a 2-kb gene region showing low recombination (Supplemental Fig. 8): The topology indicated the presence of two distantly related haplotype clusters (Fig. 6) with an estimated TMRCA of 1.76 Myr (SD = 576 Kyr, ρ = 4.99; fixed differences with chimpanzee = 34). Calculation of the TMRCA using GENETREE resulted in a comparable estimate (TMRCA = 1.53 Myr, SD = 443 Kyr, θ_{ML} = 2.5, N_e = 10,714; Supplemental Fig. 8). Such divergent haplotype clades can be expected under two different circumstances, namely balancing selection and ancient population structure. Yet, some difference exists in that symmetric balancing selection is expected to elongate the entire neutral genealogy, while the effects of ancient population structure are reflected in an increase in the genealogical time occupied by two single lineages (Takahata 1990;

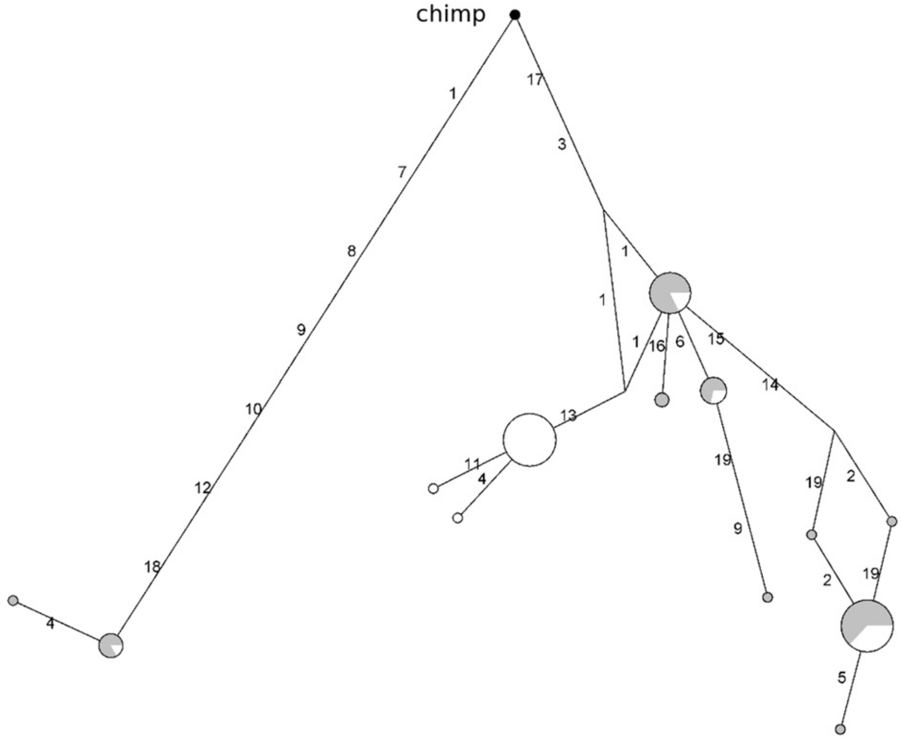

Figure 6. Genealogy of *BSG* haplotypes reconstructed through a median-joining network. The analysis corresponds to the gene region spanning nucleotides 8500–10,300. Population color codes are as follows: (white) EA; (gray) YRI.

Wall 2000). A possibility to discriminate between these scenarios is to calculate the percentage of congruent mutations, meaning those that occur on the basal branches of a genealogy (Wall 2000). When we applied this approach to the two major *BSG* clades, a percentage of congruent mutations equal to 34% was obtained; this is lower than previous estimates under a model of ancient population structure, which ranged from 42% to 45% (Barreiro et al. 2005; Garrigan et al. 2005); also, the TMRCA we estimated for the *BSG* gene is not unusual (Tishkoff and Verrelli 2003; Garrigan and Hammer 2006), while deep coalescent times are expected when ancient population subdivision is involved. The asymmetric structure of the haplotype genealogy whereby most chromosomes cluster in one clade with a relatively deep coalescent, while a minor branch is accounted for by a small number of less diverged chromosomes is difficult to interpret within a theoretical framework. Different explanations might account for the *BSG* genealogy, one appealing possibility being frequency-dependent balancing selection, accounting for the maintenance of a distantly related haplotype with a low frequency in the population. Another possibility is that different selective events have being acting on the *BSG* locus or, else, that complex demographic scenarios account for the pattern we observe nowadays.

Basigin (also known as CD147) has been involved in different biologic and pathologic processes, such as amyloid-beta production, thymocyte maturation, cellular invasion, and rheumatoid arthritis (Iacono et al. 2007). Moreover, functioning as a receptor for cyclophilin A makes CD147 a facilitator of HIV-1 infection (Pushkarsky et al. 2001). This property derives from the ability of HIV-1 to incorporate cyclophilin A into virions, a feature which is common to other viruses (Castro et al. 2003; Lin and Emerman 2006). Additional studies aimed at clarifying the evolutionary history of BSG and its role in infections might benefit from this initial description.

Conclusions

Haldane's hypothesis as formulated in 1932 posits that infectious diseases have been a major threat to human populations and have therefore exerted strong selective pressures throughout human history (Haldane 1932). A few years later he also presciently proposed that antigens constituted of protein-carbohydrates molecules account for "surprising biochemical diversity by serological tests" and possibly play a role in resistance/predisposition to pathogen infection (Haldane 1949). These lines seem to perfectly fit BGA genes, as demonstrated by both this study and previous descriptions (Saitou and Yamamoto 1997; Koda et al. 2001; Baum et al. 2002; Hamblin et al. 2002).

On the one hand, despite medical advances in treatment and prevention, infectious diseases represent a major selective pressure in humans and account for about 48% of deaths in people younger than 45 yr worldwide (Kapp 1999). On the other hand, different BGAs have been shown to act as receptors for one or more pathogens and differential disease susceptibility has been substantiated in some cases depending on BG phenotype. In this scenario, it is not surprising that BGA genes have been the target of selective pressures and associations between pathogen richness and BGA alleles can be identified.

Indeed, here we show that four BGA genes have been subjected to balancing selection (the underlying selective pressure possibly being an infectious agent) and that pathogen richness has shaped allele frequencies in 11 genes. These data, together with a previous description of non neutral evolution for *ABO* (Saitou and Yamamoto 1997; Calafell et al. 2008), *FUT2* (Koda et al. 2001), *GYPA* (Baum et al. 2002), and *DARC* (Hamblin et al. 2002), indicate that BGAs played a central role in the host–pathogen arms race during human evolutionary history.

Methods

DNA samples and sequencing

Human genomic DNA was obtained from the Coriell Institute for Medical Research. All analyzed regions were PCR amplified and directly sequenced; primer sequences are available upon request. PCR products were treated with ExoSAP-IT (USB Corporation), directly sequenced on both strands with a Big Dye Terminator sequencing Kit (v3.1 Applied Biosystem) and run on an Applied Biosystems ABI 3130 XL Genetic Analyzer (Applied Biosystem). Sequences were assembled using AutoAssembler version 1.4.0 (Applied Biosystems), inspected manually by two distinct operators, and singletons were re-amplified and resequenced.

Data retrieval and haplotype construction

Genotype data for two populations, one of African ancestry and one of Caucasian ancestry, were retrieved from the SeattleSNPs website (http://pga.mbt.washington.edu). Nucleotide positions for all analyzed genes correspond to those of SeattleSNPs, which in turn are derived from the following GenBank accession nos.: AY942196 (*BSG*), AY851161 (*CD55*), DQ074789 (*CD151*), AY937240 (*FUT2*), and AY942197 (*SLC14A1*).

Genotype data for 238 resequenced human genes were derived from the NIEHS SNPs Program website (http://egp.gs.washington.edu). In particular we selected genes that had been resequenced in populations of defined ethnicity including African American (AA), Caucasians (European ancestry, EA), Yoruba (YRI), and Asians (AS) (NIEHS panel 2). Similarly, genotype data from 304 resequenced genes were derived from the SeattleSNPs Web site. In particular, 201 and 103 genes have been resequenced across panels 1 and 2, respectively, the former containing African American and European American, the latter Yoruban and European subjects.

Haplotypes were inferred using PHASE version 2.1 (Stephens et al. 2001; Stephens and Scheet 2005), a program for reconstructing haplotypes from unrelated genotype data through a Bayesian statistical method. Haplotypes for individuals resequenced in this study are available as supplementary material (Supplemental File 3).

Linkage disequilibrium analyses were performed using Haploview (Barrett et al. 2005), and haplotype blocks were identified through an implemented method (Gabriel et al. 2002).

Data concerning HGDP-CEPH SNPs derive from a previous work (Li et al. 2008). A SNP was ascribed to a specific gene if it was located within the transcribed region or no more than 700 bp upstream from the transcription start site.

Statistical analysis

The correlation between pathogen richness and BGA allele frequencies was assessed by Kendall's rank correlation coefficient (τ), a non-parametric statistic used to measure the degree of correspondence between two rankings. The reason for using this test is that even in the presence of ties, the sampling distribution of τ satisfactorily converges to a normal distribution for values of n larger than 10 (Salkind 2007).

In order to evaluate the probability of obtaining 19 SNPs out of 262 with a τ higher than the 95th percentile, we performed 30,000 simulations. In particular, samples of 262 SNPs were extracted from the full data set by searching for each BGA SNP, one with an allele frequency matched at the 0.001 level. For each sample, the number of SNPs with a percentile rank higher than the 95th percentile (calculated over all SNPs) was counted. By this procedure, the empirical probability of obtaining 19 or more SNPs was estimated equal to 0.045. A theoretic approach can also be applied by considering that the probability to obtain n SNPs with a τ higher than the 95th percentile in a sample of 262 is Poisson-distributed with lambda = 13 (5% of 262). For such a distribution the probability of obtaining 19 or more SNPs equals 0.043.

The F_{ST} statistic (Wright 1950) estimates genetic differentiation among populations and was calculated as previously proposed (Hudson et al. 1992). In order to identify gene regions showing extreme F_{ST}-values, sliding windows of 5 kb moving along BGA genes with a step of 150 bp were used; the same procedure was applied to all genes resequenced by the SeattleSNPs program. Values deriving from sliding windows obtained from all genes resequenced in panels 1 and 2 were used to identify the 2.5th and 97.5th percentiles that represented the threshold to define unusually high or low F_{ST}-values in BGA genes. It is worth noting that negative F_{ST} should be interpreted as 0 and the 2.5th percentile value of F_{ST} from SeattleSNPs gene sliding windows resulted extremely close to 0. Therefore, BGA windows displaying an F_{ST}-value negative or equal to 0 were considered to display exceedingly low population differentiation. In order to evaluate the probability of obtaining 8.3% of windows showing F_{ST}-values, either below the 2.5th or above the 97.5th percentiles, we used a simulation-based approach. In particular, 17 genes were randomly selected from the SeattleSNPs database and for each group the fraction of sliding windows showing exceedingly low or high values was counted. Ten thousand simulations were performed and the probability of obtaining a fraction of outliers equal to or higher than 8.3% was calculated.

Tajima's D (Tajima 1989), Fu and Li's D^* and F^* (Fu and Li 1993) statistics, as well as diversity parameters θ_W (Watterson 1975) and π (Nei and Li 1979) and Fay and Wu's H (Fay and Wu 2000) were calculated using *libsequence* (Thornton 2003), a C++ class library providing an object-oriented framework for the analysis of molecular population genetic data. Calibrated coalescent simulations were performed using the *cosi* package (Schaffner et al. 2005) and its best-fit parameters for YRI, AA, EA, and AS populations with 10,000 iterations. As a further control, summary statistics were calculated for 5 kb windows deriving from NIEHS genes and the values obtained for BGA gene regions compared to their distribution. In particular, for each gene a 5 kb region was randomly selected; the only requirement was that it did not contain any long (>500 bp) resequencing gap; if the gene did not fulfill this requirement it was discarded, as were 5 kb regions displaying less than five SNPs. The numbers of analyzed windows for AA, YRI, EA, and AS were 209, 203, 177, and 172, respectively. The same procedure was applied to SeattleSNPs genes and a total of 103, 201, and 298 windows were obtained for YRI, AA, and subjects with European ancestry, respectively.

The maximum-likelihood-ratio HKA test was performed using the MLHKA software (Wright and Charlesworth 2004) using multilocus data of 16 genes and *Pan troglodytes* (NCBI panTro2) as an outgroup. The 16 reference genes were randomly selected among NIEHS loci shorter than 20 kb that have been resequenced in the four populations (YRI, AA, EA, and AS; panel 2); the only criterion was that Tajima's D did not suggest the action of natural selection (i.e., D_T is higher than the 2.5th and lower than the 97.5th percentiles in the distribution of NIEHS genes; see Supplemental Table 3). The reference set was accounted for by the following genes: *VNN3, PLA2G2D, MB, MAD2L2, HRAS, CYP17A1, ATOX1, BNIP3, CDC20, NGB, TUBA1, MT3, NUDT1, PRDX5, RETN*, and *JUND*.

We evaluated the likelihood of the model under two different assumptions: that all loci evolved neutrally and that only the region under analysis was subjected to natural selection; statistical significance was assessed by a likelihood ratio test. We used a chain length (the number of cycles of the Markov chain) of 2×10^5 and, as suggested by Wright and Charlesworth (2004), we ran the program several times with different seeds to ensure stability of results.

Median-joining networks to infer haplotype genealogy was constructed using NETWORK 4.5 (Bandelt et al. 1999). Estimate of the time to the most common ancestor (TMRCA) was obtained using a phylogeny based approach implemented in NETWORK using a mutation rate based on the number of fixed differences between human and chimpanzee or orangutan and assuming a separation time from humans of 6 Myr and 13 Myr ago, respectively (Glazko and Nei 2003). In all cases, a second TMRCA estimate derived from application of a maximum-likelihood coalescent method implemented in GENETREE (Griffiths and Tavare 1994, 1995). Again, the mutation rate μ was obtained on the basis of the divergence between human and a primate, assuming a generation time of 25 yr. In using this μ and the estimated maximum likelihood θ (θ_{ML}), we estimated the effective population size parameter (N_e). With these assumptions, the coalescence time, scaled in $2N_e$ units, was converted into years. For the coalescence process, 10^6 simulations were performed. All calculations were performed in the R environment (www.r-project.org).

Environmental variables

Pathogen absence/presence matrices for the 21 countries where HGDP-CEPH populations are located were derived from the Gideon database (http://www.gideononline.com) following previous indications (Prugnolle et al. 2005). Briefly, only species that are transmitted in the countries were included, meaning that cases of transmission due to tourism and immigration were not taken into account; also, species that have recently been eradicated as a result, for example, of vaccination campaigns, were recorded as present in the matrix. It should be noted that the final number of different pathogen species per country differ from those calculated by Prugnolle et al. (2005), since these authors only took into account intracellular disease agents. Precipitation rate and mean temperature were derived for the geographic coordinates corresponding to HGDP-CEPH populations from the NCEP/NCAR database (Kistler et al. 2001).

Sequence annotation

Data concerning DNase I hypersensitive sites in CD4+ T cells derive from a previous work (Boyle et al. 2008) and were retrieved from the UCSC annotation tables (http://genome.ucsc.edu, Duke DNase I HS track). MicroRNA binding sites were identified through the dedicated utility at miRBase, which relies on the miRanda algorithm (John et al. 2004) and requires a target site to be conserved in at least two species. Functional elements in 3′UTR were searched for using UTRscan (Pesole and Liuni 1999).

Acknowledgments

We thank Dr. Roberto Giorda for helpful comments and discussion about the manuscript. M.S. and R.C. are part of the Doctorate School of Molecular Medicine, University of Milan.

References

Akey, J.M., Swanson, W.J., Madeoy, J., Eberle, M., and Shriver, M.D. 2006. TRPV6 exhibits unusual patterns of polymorphism and divergence in worldwide populations. *Hum. Mol. Genet.* **15:** 2106–2113.

Ali, S., Niang, M.A., N'doye, I., Critchlow, C.W., Hawes, S.E., Hill, A.V., and Kiviat, N.B. 2000. Secretor polymorphism and human immunodeficiency virus infection in Senegalese women. *J. Infect. Dis.* **181:** 737–739.

Bandelt, H.J., Forster, P., and Rohl, A. 1999. Median-joining networks for inferring intraspecific phylogenies. *Mol. Biol. Evol.* **16:** 37–48.

Barreiro, L.B., Patin, E., Neyrolles, O., Cann, H.M., Gicquel, B., and Quintana-Murci, L. 2005. The heritage of pathogen pressures and ancient demography in the human innate-immunity CD209/CD209L region. *Am. J. Hum. Genet.* **77:** 869–886.

Barrett, J.C., Fry, B., Maller, J., and Daly, M.J. 2005. Haploview: Analysis and visualization of LD and haplotype maps. *Bioinformatics* **21:** 263–265.

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

Blackwell, C.C., Weir, D.M., James, V.S., Todd, W.T., Banatvala, N., Chaudhuri, A.K., Gray, H.G., Thomson, E.J., and Fallon, R.J. 1990. Secretor status, smoking and carriage of *Neisseria meningitidis*. *Epidemiol. Infect.* **104:** 203–209.

Blumenfeld, O.O. and Patnaik, S.K. 2004. Allelic genes of blood group antigens: A source of human mutations and cSNPs documented in the Blood Group Antigen Gene Mutation Database. *Hum. Mutat.* **23:** 8–16.

Boyle, A.P., Davis, S., Shulha, H.P., Meltzer, P., Margulies, E.H., Weng, Z., Furey, T.S., and Crawford, G.E. 2008. High-resolution mapping and characterization of open chromatin across the genome. *Cell* **132:** 311–322.

Bubb, K.L., Bovee, D., Buckley, D., Haugen, E., Kibukawa, M., Paddock, M., Palmieri, A., Subramanian, S., Zhou, Y., Kaul, R., et al. 2006. Scan of human genome reveals no new loci under ancient balancing selection. *Genetics* **173:** 2165–2177.

Calafell, F., Roubinet, F., Ramírez-Soriano, A., Saitou, N., Bertranpetit, J., and Blancher, A. 2008. Evolutionary dynamics of the human ABO gene. *Hum. Genet.* **124:** 123–135.

Casanova, J.L. and Abel, L. 2007. Human genetics of infectious diseases: A unified theory. *EMBO J.* **26:** 915–922.

Castro, A.P., Carvalho, T.M., Moussatche, N., and Damaso, C.R. 2003. Redistribution of cyclophilin A to viral factories during vaccinia virus infection and its incorporation into mature particles. *J. Virol.* **77:** 9052–9068.

Cereb, N., Hughes, A.L., and Yang, S.Y. 1997. Locus-specific conservation of the HLA class I introns by intra-locus homogenization. *Immunogenetics* **47:** 30–36.

Chaturvedi, P., Warren, C.D., Altaye, M., Morrow, A.L., Ruiz-Palacios, G., Pickering, L.K., and Newburg, D.S. 2001. Fucosylated human milk oligosaccharides vary between individuals and over the course of lactation. *Glycobiology* **11:** 365–372.

Chimpanzee Sequencing and Analysis Consortium. 2005. Initial sequence of the chimpanzee genome and comparison with the human genome. *Nature* **437:** 69–87.

Clarkson, N.A., Kaufman, R., Lublin, D.M., Ward, T., Pipkin, P.A., Minor, P.D., Evans, D.J., and Almond, J.W. 1995. Characterization of the echovirus 7 receptor: Domains of CD55 critical for virus binding. *J. Virol.* **69:** 5497–5501.

Costache, M., Apoil, P.A., Cailleau, A., Elmgren, A., Larson, G., Henry, S., Blancher, A., Iordachescu, D., Oriol, R., and Mollicone, R. 1997. Evolution of fucosyltransferase genes in vertebrates. *J. Biol. Chem.* **272:** 29721–29728.

Cowin, A.J., Adams, D., Geary, S.M., Wright, M.D., Jones, J.C., and Ashman, L.K. 2006. Wound healing is defective in mice lacking tetraspanin CD151. *J. Invest. Dermatol.* **126:** 680–689.

Cywes, C., Stamenkovic, I., and Wessels, M.R. 2000. CD44 as a receptor for colonization of the pharynx by group A Streptococcus. *J. Clin. Invest.* **106:** 995–1002.

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

Felsenfeld, G. and Groudine, M. 2003. Controlling the double helix. *Nature* **421:** 448–453.

Fitter, S., Tetaz, T.J., Berndt, M.C., and Ashman, L.K. 1995. Molecular cloning of cDNA encoding a novel platelet-endothelial cell tetra-span antigen, PETA-3. *Blood* **86:** 1348–1355.

Fu, Y.X. and Li, W.H. 1993. Statistical tests of neutrality of mutations. *Genetics* **133:** 693–709.

Fu, Y., Liu, Z., Lin, J., Chen, W., Jia, Z., Pan, D., and Xu, A. 2003. Extensive polymorphism and different evolutionary patterns of intron 2 were identified in the *HLA-DQB1* gene. *Immunogenetics* **54:** 761–766.

Gabriel, S.B., Schaffner, S.F., Nguyen, H., Moore, J.M., Roy, J., Blumenstiel, B., Higgins, J., DeFelice, M., Lochner, A., Faggart, M., et al. 2002. The structure of haplotype blocks in the human genome. *Science* **296:** 2225–2229.

Gagneux, P. and Varki, A. 1999. Evolutionary considerations in relating oligosaccharide diversity to biological function. *Glycobiology* **9:** 747–755.

Garrigan, D. and Hammer, M.F. 2006. Reconstructing human origins in the genomic era. *Nat. Rev. Genet.* **7:** 669–680.

Garrigan, D., Mobasher, Z., Kingan, S.B., Wilder, J.A., and Hammer, M.F. 2005. Deep haplotype divergence and long-range linkage disequilibrium at xp21.1 provide evidence that humans descend from a structured ancestral population. *Genetics* **170:** 1849–1856.

Glazko, G.V. and Nei, M. 2003. Estimation of divergence times for major lineages of primate species. *Mol. Biol. Evol.* **20:** 424–434.

Greenwell, P. 1997. Blood group antigens: Molecules seeking a function? *Glycoconj. J.* **14:** 159–173.

Griffiths, R.C. and Tavare, S. 1994. Sampling theory for neutral alleles in a varying environment. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **344:** 403–410.

Griffiths, R.C. and Tavare, S. 1995. Unrooted genealogical tree probabilities in the infinitely-many-sites model. *Math. Biosci.* **127:** 77–98.

Gross, D.S. and Garrard, W.T. 1988. Nuclease hypersensitive sites in chromatin. *Annu. Rev. Biochem.* **57:** 159–197.

Guernier, V., Hochberg, M.E., and Guegan, J.F. 2004. Ecology drives the worldwide distribution of human diseases. *PLoS Biol.* **2:** e141. doi: 10.1371/journal.pbio.0020141.

Haldane, J.B.S. 1932. *The causes of evolution*. Longmans, Green & Co., London, UK.

Haldane, J.B.S. 1949. Disease and evolution. Symposium sui fattori ecologici e genetici della speciazione negli animali. In *Selected genetic papers of J.B.S. Haldane* (Anonymous), pp. 325–334. Garland Publishing Inc., New York/London.

Hamblin, M.T., Thompson, E.E., and Di Rienzo, A. 2002. Complex signatures of natural selection at the Duffy blood group locus. *Am. J. Hum. Genet.* **70:** 369–383.

Hancock, A.M., Witonsky, D.B., Gordon, A.S., Eshel, G., Pritchard, J.K., Coop, G., and Di Rienzo, A. 2008. Adaptations to climate in candidate genes for common metabolic disorders. *PLoS Genet.* **4:** e32. doi: 10.1371/journal.pgen.0040032.

Handley, L.J., Manica, A., Goudet, J., and Balloux, F. 2007. Going the distance: Human population genetics in a clinal world. *Trends Genet.* **23:** 432–439.

Haverkorn, M.J. and Goslings, W.R. 1969. Streptococci, ABO blood groups, and secretor status. *Am. J. Hum. Genet.* **21:** 360–375.

Hill, A.V. 2006. Aspects of genetic susceptibility to human infectious diseases. *Annu. Rev. Genet.* **40:** 469–486.

Ho, S.H., Martin, F., Higginbottom, A., Partridge, L.J., Parthasarathy, V., Moseley, G.W., Lopez, P., Cheng-Mayer, C., and Monk, P.N. 2006. Recombinant extracellular domains of tetraspanin proteins are potent inhibitors of the infection of macrophages by human immunodeficiency virus type 1. *J. Virol.* **80:** 6487–6496.

Hudson, R.R., Kreitman, M., and Aguade, M. 1987. A test of neutral molecular evolution based on nucleotide data. *Genetics* **116:** 153–159.

Hudson, R.R., Slatkin, M., and Maddison, W.P. 1992. Estimation of levels of gene flow from DNA sequence data. *Genetics* **132:** 583–589.

Hurd, E.A. and Domino, S.E. 2004. Increased susceptibility of secretor factor gene *Fut2*-null mice to experimental vaginal candidiasis. *Infect. Immun.* **72:** 4279–4281.

Iacono, K.T., Brown, A.L., Greene, M.I., and Saouaf, S.J. 2007. CD147 immunoglobulin superfamily receptor function and role in pathology. *Exp. Mol. Pathol.* **83:** 283–295.

John, B., Enright, A.J., Aravin, A., Tuschl, T., Sander, C., and Marks, D.S. 2004. Human microRNA targets. *PLoS Biol.* **2:** e363. doi: 10.1371/journal.pbio.0020363.

Kapp, C. 1999. WHO warns of microbial threat. *Lancet* **353:** 2222. doi: 1016/S0140-6736(05)76281-4.

Karamatic Crew, V., Burton, N., Kagan, A., Green, C.A., Levene, C., Flinter, F., Brady, R.L., Daniels, G., and Anstee, D.J. 2004. CD151, the first member of the tetraspanin (TM4) superfamily detected on erythrocytes, is essential for the correct assembly of human basement membranes in kidney and skin. *Blood* **104:** 2217–2223.

Kaul, A., Nagamani, M., and Nowicki, B. 1995. Decreased expression of endometrial decay accelerating factor (DAF), a complement regulatory protein, in patients with luteal phase defect. *Am. J. Reprod. Immunol.* **34:** 236–240.

Kelly, R.J., Rouquier, S., Giorgi, D., Lennon, G.G., and Lowe, J.B. 1995. Sequence and expression of a candidate for the human secretor blood group alpha(1,2)fucosyltransferase gene (FUT2). Homozygosity for an enzyme-inactivating nonsense mutation commonly correlates with the non-secretor phenotype. *J. Biol. Chem.* **270:** 4640–4649.

Kendall, M.G. 1976. *Rank correlation methods*. Griffin, London.

Kimura, M. 1983. *The neutral theory of molecular evolution*. Cambridge University Press, Cambridge.

Kistler, R., Kalnay, E., Collins, W., Saha, S., White, G., Woollen, J., Chelliah, M., Ebisuzaki, W., Kanamitsu, M., Kousky, V., et al. 2001. The NCEP-NCAR 50-year reanalysis: Monthly means CD-ROM and documentation. *Bull. Am. Meteorol. Soc.* **82:** 247–268.

Koda, Y., Soejima, M., Liu, Y., and Kimura, H. 1996. Molecular basis for secretor type alpha(1,2)-fucosyltransferase gene deficiency in a Japanese population: A fusion gene generated by unequal crossover responsible for the enzyme deficiency. *Am. J. Hum. Genet.* **59:** 343–350.

Koda, Y., Tachida, H., Pang, H., Liu, Y., Soejima, M., Ghaderi, A.A., Takenaka, O., and Kimura, H. 2001. Contrasting patterns of polymorphisms at the ABO-secretor gene (FUT2) and plasma alpha(1,3)fucosyltransferase gene (FUT6) in human populations. *Genetics* **158:** 747–756.

Leemans, J.C., Florquin, S., Heikens, M., Pals, S.T., van der Neut, R., and Van Der Poll, T. 2003. CD44 is a macrophage binding site for Mycobacterium tuberculosis that mediates macrophage recruitment and protective immunity against tuberculosis. *J. Clin. Invest.* **111:** 681–689.

Li, J.Z., Absher, D.M., Tang, H., Southwick, A.M., Casto, A.M., Ramachandran, S., Cann, H.M., Barsh, G.S., Feldman, M., Cavalli-Sforza, L.L., et al. 2008. Worldwide human relationships inferred from genome-wide patterns of variation. *Science* **319:** 1100–1104.

Lin, T.Y. and Emerman, M. 2006. Cyclophilin A interacts with diverse lentiviral capsids. *Retrovirology* **3:** 70. doi: 10.1186/1742-4690-3-70.

Linden, S., Mahdavi, J., Semino-Mora, C., Olsen, C., Carlstedt, I., Boren, T., and Dubois, A. 2008. Role of ABO secretor status in mucosal innate immunity and *H. pylori* infection. *PLoS Pathog.* **4:** e2. doi: 10.1371/journal.ppat.004002.

Lindesmith, L., Moe, C., Marionneau, S., Ruvoen, N., Jiang, X., Lindblad, L., Stewart, P., LePendu, J., and Baric, R. 2003. Human susceptibility and resistance to Norwalk virus infection. *Nat. Med.* **9:** 548–553.

Liu, Y., Koda, Y., Soejima, M., Pang, H., Schlaphoff, T., du Toit, E.D., and Kimura, H. 1998. Extensive polymorphism of the *FUT2* gene in an African (Xhosa) population of South Africa. *Hum. Genet.* **103:** 204–210.

Liu, Y., Promeneur, D., Rojek, A., Kumar, N., Frokiaer, J., Nielsen, S., King, L.S., Agre, P., and Carbrey, J.M. 2007. Aquaporin 9 is the major pathway for glycerol uptake by mouse erythrocytes, with implications for malarial virulence. *Proc. Natl. Acad. Sci.* **104:** 12560–12564.

Loffler, S., Lottspeich, F., Lanza, F., Azorsa, D.O., ter Meulen, V., and Schneider-Schaulies, J. 1997. CD9, a tetraspan transmembrane protein, renders cells susceptible to canine distemper virus. *J. Virol.* **71:** 42–49.

Lui, W.O., Pourmand, N., Patterson, B.K., and Fire, A. 2007. Patterns of known and novel small RNAs in human cervical cancer. *Cancer Res.* **67:** 6031–6043.

Moulds, J.M. and Moulds, J.J. 2000. Blood group associations with parasites, bacteria, and viruses. *Transfus. Med. Rev.* **14:** 302–311.

Moulds, J.M., Nowicki, S., Moulds, J.J., and Nowicki, B.J. 1996. Human blood groups: Incidental receptors for viruses and bacteria. *Transfusion* **36:** 362–374.

Nei, M. and Li, W.H. 1979. Mathematical model for studying genetic variation in terms of restriction endonucleases. *Proc. Natl. Acad. Sci.* **76:** 5269–5273.

Newburg, D.S., Ruiz-Palacios, G.M., and Morrow, A.L. 2005. Human milk glycans protect infants against enteric pathogens. *Annu. Rev. Nutr.* **25:** 37–58.

Nicholson-Weller, A. and Wang, C.E. 1994. Structure and function of decay accelerating factor CD55. *J. Lab. Clin. Med.* **123:** 485–491.

Nowicki, B., Hart, A., Coyne, K.E., Lublin, D.M., and Nowicki, S. 1993. Short consensus repeat-3 domain of recombinant decay-accelerating factor is recognized by *Escherichia coli* recombinant Dr adhesin in a model of a cell-cell interaction. *J. Exp. Med.* **178:** 2115–2121.

Pesole, G. and Liuni, S. 1999. Internet resources for the functional analysis of 5′ and 3′ untranslated regions of eukaryotic mRNAs. *Trends Genet.* **15:** 378. doi: 10.1016/S0168-9525(99)01795-3.

Pileri, P., Uematsu, Y., Campagnoli, S., Galli, G., Falugi, F., Petracca, R., Weiner, A.J., Houghton, M., Rosa, D., Grandi, G., et al. 1998. Binding of hepatitis C virus to CD81. *Science* **282:** 938–941.

Prugnolle, F., Manica, A., Charpentier, M., Guegan, J.F., Guernier, V., and Balloux, F. 2005. Pathogen-driven selection and worldwide HLA class I diversity. *Curr. Biol.* **15:** 1022–1027.

Przeworski, M. 2002. The signature of positive selection at randomly chosen loci. *Genetics* **160:** 1179–1189.

Pushkarsky, T., Zybarth, G., Dubrovsky, L., Yurchenko, V., Tang, H., Guo, H., Toole, B., Sherry, B., and Bukrinsky, M. 2001. CD147 facilitates HIV-1 infection by interacting with virus-associated cyclophilin A. *Proc. Natl. Acad. Sci.* **98:** 6360–6365.

Raza, M.W., Blackwell, C.C., Molyneaux, P., James, V.S., Ogilvie, M.M., Inglis, J.M., and Weir, D.M. 1991. Association between secretor status and respiratory viral illness. *BMJ* **303:** 815–818.

Reid, M.E. and Lomas-Frances, C. 1997. *The blood group antigen facts book*. Academic, San Diego.

Roudier, N., Bailly, P., Gane, P., Lucien, N., Gobin, R., Cartron, J.P., and Ripoche, P. 2002. Erythroid expression and oligomeric state of the AQP3 protein. *J. Biol. Chem.* **277:** 7664–7669.

Rouschop, K.M., Sylva, M., Teske, G.J., Hoedemaeker, I., Pals, S.T., Weening, J.J., van der Poll, T., and Florquin, S. 2006. Urothelial CD44 facilitates *Escherichia coli* infection of the murine urinary tract. *J. Immunol.* **177:** 7225–7232.

Ruiz-Palacios, G.M., Cervantes, L.E., Ramos, P., Chavez-Munguia, B., and Newburg, D.S. 2003. *Campylobacter jejuni* binds intestinal H(O) antigen (Fucα1, 2Galβ1, 4GlcNAc), and fucosyloligosaccharides of human milk inhibit its binding and infection. *J. Biol. Chem.* **278:** 14112–14120.

Sachs, N., Kreft, M., van den Bergh Weerman, M.A., Beynon, A.J., Peters, T.A., Weening, J.J., and Sonnenberg, A. 2006. Kidney failure in mice lacking the tetraspanin CD151. *J. Cell Biol.* **175:** 33–39.

Saitou, N. and Yamamoto, F. 1997. Evolution of primate ABO blood group genes and their homologous genes. *Mol. Biol. Evol.* **14:** 399–411.

Salkind, N.J. 2007. *Encyclopedia of measurement and statistics*. Sage Publications, Thousand Oaks, CA.

Sands, J.M., Gargus, J.J., Frohlich, O., Gunn, R.B., and Kokko, J.P. 1992. Urinary concentrating ability in patients with Jk(a-b-) blood type who lack carrier-mediated urea transport. *J. Am. Soc. Nephrol.* **2:** 1689–1696.

Schaeffer, A.J., Rajan, N., Cao, Q., Anderson, B.E., Pruden, D.L., Sensibar, J., and Duncan, J.L. 2001. Host pathogenesis in urinary tract infections. *Int. J. Antimicrob. Agents* **17:** 245–251.

Schaffner, S.F., Foo, C., Gabriel, S., Reich, D., Daly, M.J., and Altshuler, D. 2005. Calibrating a coalescent simulation of human genome sequence variation. *Genome Res.* **15:** 1576–1583.

Sendide, K., Deghmane, A.E., Reyrat, J.M., Talal, A., and Hmama, Z. 2004. *Mycobacterium bovis* BCG urease attenuates major histocompatibility complex class II trafficking to the macrophage cell surface. *Infect. Immun.* **72:** 4200–4209.

Shafren, D.R., Bates, R.C., Agrez, M.V., Herd, R.L., Burns, G.F., and Barry, R.D. 1995. Coxsackieviruses B1, B3, and B5 use decay accelerating factor as a receptor for cell attachment. *J. Virol.* **69:** 3873–3877.

Shanmukhappa, K., Kim, J.K., and Kapil, S. 2007. Role of CD151, A tetraspanin, in porcine reproductive and respiratory syndrome virus infection. *Virol. J.* **4:** 62. doi: 10.1186/1743-422X-4-62.

Soejima, M. and Koda, Y. 2008. Distinct single nucleotide polymorphism pattern at the *FUT2* promoter among human populations. *Ann. Hematol.* **87:** 19–25.

Sood, R., Zehnder, J.L., Druzin, M.L., and Brown, P.O. 2006. Gene expression patterns in human placenta. *Proc. Natl. Acad. Sci.* **103:** 5478–5483.

Stephens, M. and Scheet, P. 2005. Accounting for decay of linkage disequilibrium in haplotype inference and missing-data imputation. *Am. J. Hum. Genet.* **76:** 449–462.

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.

Tajima, F. 1989. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics* **123:** 585–595.

Takahata, N. 1990. A simple genealogical structure of strongly balanced allelic lines and *trans*-species evolution of polymorphism. *Proc. Natl. Acad. Sci.* **87:** 2419–2423.

Takahata, N. and Satta, Y. 1998. Footprints of intragenic recombination at HLA loci. *Immunogenetics* **47:** 430–441.

Thompson, E.E., Kuttab-Boulos, H., Witonsky, D., Yang, L., Roe, B.A., and Di Rienzo, A. 2004. CYP3A variation and the evolution of salt-sensitivity variants. *Am. J. Hum. Genet.* **75:** 1059–1069.

Thornton, K. 2003. libsequence: A C++ class library for evolutionary genetic analysis. *Bioinformatics* **19:** 2325–2327.

Tishkoff, S.A. and Verrelli, B.C. 2003. Patterns of human genetic diversity: Implications for human evolutionary history and disease. *Annu. Rev. Genomics Hum. Genet.* **4:** 293–340.

von Lindern, J.J., Rojo, D., Grovit-Ferbas, K., Yeramian, C., Deng, C., Herbein, G., Ferguson, M.R., Pappas, T.C., Decker, J.M., Singh, A., et al. 2003. Potential role for CD63 in CCR5-mediated human immunodeficiency virus type 1 infection of macrophages. *J. Virol.* **77:** 3624–3633.

Wall, J.D. 2000. Detecting ancient admixture in humans using sequence polymorphism data. *Genetics* **154:** 1271–1279.

Watterson, G.A. 1975. On the number of segregating sites in genetical models without recombination. *Theor. Popul. Biol.* **7:** 256–276.

Willett, B.J., Hosie, M.J., Jarrett, O., and Neil, J.C. 1994. Identification of a putative cellular receptor for feline immunodeficiency virus as the feline homologue of CD9. *Immunology* **81:** 228–233.

Wiuf, C., Zhao, K., Innan, H., and Nordborg, M. 2004. The probability and chromosomal extent of *trans*-specific polymorphism. *Genetics* **168:** 2363–2372.

Wright, S. 1950. Genetical structure of populations. *Nature* **166:** 247–249.

Wright, S.I. and Charlesworth, B. 2004. The HKA test revisited: A maximum-likelihood-ratio test of the standard neutral model. *Genetics* **168:** 1071–1076.

Wright, M.D., Geary, S.M., Fitter, S., Moseley, G.W., Lau, L.M., Sheng, K.C., Apostolopoulos, V., Stanley, E.G., Jackson, D.E., and Ashman, L.K. 2004. Characterization of mice lacking the tetraspanin superfamily member CD151. *Mol. Cell. Biol.* **24:** 5978–5988.

Yang, B., Bankir, L., Gillespie, A., Epstein, C.J., and Verkman, A.S. 2002. Urea-selective concentrating defect in transgenic mice lacking urea transporter UT-B. *J. Biol. Chem.* **277:** 10633–10637.

Young, S.L., Lessey, B.A., Fritz, M.A., Meyer, W.R., Murray, M.J., Speckman, P.L., and Nowicki, B.J. 2002. In vivo and in vitro evidence suggest that HB-EGF regulates endometrial expression of human decay-accelerating factor. *J. Clin. Endocrinol. Metab.* **87:** 1368–1375.

Young, J.H., Chang, Y.P., Kim, J.D., Chretien, J.P., Klag, M.J., Levine, M.A., Ruff, C.B., Wang, N.Y., and Chakravarti, A. 2005. Differential susceptibility to hypertension is due to selection during the out-of-Africa expansion. *PLoS Genet.* **1:** e82. doi: 10.1371/journal.pgen.0010082.

Received June 30, 2008; accepted in revised form November 4, 2008.