Am. J. Hum. Genet. 66:1669–1679, 2000

Detection of the Signature of Natural Selection in Humans: Evidence from the Duffy Blood Group Locus

Martha T. Hamblin and Anna Di Rienzo

Department of Human Genetics, University of Chicago, Chicago

The Duffy blood group locus, which encodes a chemokine receptor, is characterized by three alleles—*FY*A*, *FY*B*, and *FY*O*. The frequency of the *FY*O* allele, which corresponds to the absence of Fy antigen on red blood cells, is at or near fixation in most sub-Saharan African populations but is very rare outside Africa. The F_{ST} value for the *FY*O* allele is the highest observed for any allele in humans, providing strong evidence for the action of natural selection at this locus. Homozygosity for the *FY*O* allele confers complete resistance to vivax malaria, suggesting that this allele has been the target of selection by *Plasmodium vivax* or some other infectious agent. To characterize the signature of directional selection at this locus, we surveyed DNA sequence variation, both in a 1.9-kb region centered on the *FY*O* mutation site and in a 1-kb region 5–6 kb away from it, in 17 Italians and in a total of 24 individuals from five sub-Saharan African populations. The level of variation across both regions is two- to threefold lower in the Africans than in the Italians. As a result, the pooled African sample shows a significant departure from the neutral expectation for the number of segregating sites, whereas the Italian sample does not. The *FY*O* allele occurs on two major haplotypes in three of the five African populations. This finding could be due to recombination, recurrent mutation, population structure, and/or mutation accumulation and drift. Although we are unable to distinguish among these alternative hypotheses, it is likely that the two major haplotypes originated prior to selection on the *FY*O* mutation.

Introduction

The Duffy blood group locus is characterized by three main alleles—*FY*A* and *FY*B*, which differ by a single amino acid, and *FY*O*, which corresponds to the Fy(a−b−) serological phenotype (i.e., the absence of Fy antigen). In Asia and the Pacific the *FY*A* allele is at high frequency, whereas in Europe and the Americas the *FY*A* and *FY*B* alleles are at intermediate frequencies. The *FY*O* allele is at or near fixation in most sub-Saharan African populations but is very rare outside Africa. As a result, the level of interpopulation differentiation of the *FY*O* allele, as measured by the F_{ST} statistic, is the highest observed for any allele in humans (Cavalli-Sforza et al. 1994). These findings strongly suggest that the observed pattern of allele frequencies at this locus has been driven by positive natural selection. This hypothesis is further supported by the observation that individuals homozygous for the *FY*O* allele are completely resistant to vivax malaria (Miller et al. 1976).

The three major Duffy alleles were originally defined serologically and recently have been characterized at the molecular level. By homology, the Duffy gene product is a member of the family of chemokine receptors. It is located in the membrane of red blood cells (RBCs), in addition to several other cell types, and binds chemokines of both the C-C and C-X-C families, but its physiological function is still unclear (Hadley and Peiper 1997). *Plasmodium vivax* requires that the Duffy antigen receptor for chemokines (DARC) be present on the RBC surface, in order to invade the cells and cause disease. Furthermore, the *FY*O* mutation has been identified as a single base change in a GATA1 regulatory element, resulting in the elimination of transcription of DARC mRNA in RBCs but not in other cell types (Tournamille et al. 1995). Thus, the connection between the Fy(a−b−) serological phenotype and resistance to vivax malaria is direct and well understood, providing a plausible hypothesis for the cause of the selective fixation of the *FY*O* allele in much of Africa.

Consideration of natural selection is critical for an understanding of the levels and patterns of variation across the genome. Figure 1 shows an idealized genealogy under directional selection: a new mutation goes deterministically to fixation without recombination, and all branches except the one carrying the mutation are lost, resulting in a loss of neutral variation linked to the selected site. This model for the effect of directional selection on patterns of variation linked to an

Received January 31, 2000; accepted for publication February 28, 2000; electronically published April 12, 2000.

Address for correspondence and reprints: Dr. Anna Di Rienzo, Department of Human Genetics, University of Chicago, 920 East 58th Street, Chicago, IL 60637. E-mail: dirienzo@genetics.uchicago.edu

0002-9297/2000/6605-0021$02.00

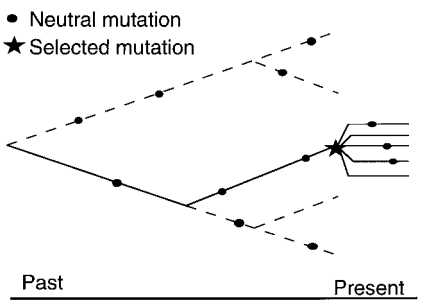

Figure 1 Genealogy of a genomic region affected by a simple selective sweep. The dashed lines indicate lineages that are lost during the sweep. Note the difference in shape of the genealogy before and after the sweep.

advantageous mutation is usually referred to as a "selective sweep." This process was first modeled by Maynard Smith and Haigh (1974), who showed that the extent of the loss of variation is a function of the strength of selection, which determines the time to fixation, and the rate of recombination. After the sweep is complete, new mutations start accumulating, but, because the genealogy is "star-shaped," there will be a preponderance of low-frequency variants (Hudson 1990). Thus, this model of directional selection makes two distinct predictions about patterns of variability: one prediction is that there will be a reduction of polymorphism, and the other prediction is that there is a skew in the spectrum of allele frequencies.

Although it is clear that directional selection can have a marked impact on patterns of sequence variation, little empirical work has been devoted to the characterization of that impact in humans. Because directional selection is a fundamental process in speciation and adaptation, the ability to identify targets of directional selection in humans is critical to an understanding of the evolution of our species. Many loci involved in the transition to anatomically modern humans are expected to have experienced directional selection on a specieswide basis at the time of that transition, whereas many other adaptations, such as those due to spatial and temporal variation in selection pressures (e.g., climate, exposure to pathogens, and diet), may have been restricted to particular populations.

The few published studies of intraspecific sequence variation for which there has been a prior hypothesis of directional selection have considered several loci putatively involved in sexual selection in the fruit fly (Aguade 1998; Tsaur et al. 1998), abalone (Metz et al. 1998), sea urchin (Metz and Palumbi 1996), and house mouse (Karn and Nachman 1999), as well as the *teosinte branched1* locus in maize (Wang et al. 1999), which has been under artificial selection during domestication. In none of these studies, however, had the actual mutation(s) under selection been identified. Therefore, the Duffy-blood-group locus provides a unique opportunity to characterize the impact of selection on patterns of sequence variation, as a function of the distance from a targeted mutation. In this study, we focus on variation around the *FY*O* mutation in sub-Saharan African populations in which the *FY*O* allele is virtually fixed, and we compare the empirical data versus theoretical expectations based on the simple selective-sweep model.

Subjects and Methods

Population Samples

Sequence variation was surveyed in five population samples from sub-Saharan Africa, comprising 5 Beti and 5 Hausa from Yaounde (Cameroon), 5 Mbuti Pygmies from the Central African Republic, 5 Mandinka from the Bamako area (The Gambia), and 4 Luo from Saradidi (Kenya); in addition, we sequenced a sample of 17 individuals from central Italy. This study was approved by the Institutional Review Board of the University of Chicago.

Sequence Determination

All primers were designed on the basis of the sequence of bacterial-artificial-chromosome clone bk134P22 (GenBank accession number AL035403), which carries an *FY*A* allele. All nucleotide positions mentioned in this article refer to this sequence. In particular, the number given in parentheses after a primer sequence is the position of the 5′ nucleotide of the primer.

PCR products were prepared for sequence analysis, either by use of the QIAquick PCR purification kit (Qiagen) or by treatment with a combination of shrimp alkaline phosphatase and exonuclease I (USB). Dye-terminator sequencing was performed with the ABI BigDye kit, and products were analyzed on an ABI 377 or 3700 automated sequencer. Chromatograms were imported into SEQUENCHER 3.1.1 (Gene Codes), for assembly into contigs and identification of polymorphic sites. Diploid sequence was determined, on both strands, for all individuals.

Analysis of the Duffy Gene Region

In most cases, a 3.2-kb product was amplified by use of forward primer 5′-GCCCTTCCTTTCCAGAGAGT-3′ (nucleotide 74119) and reverse primer 5′-TAAGAAACCACCCGCTTCAC-3′ (nucleotide 77292). For some of the individuals, we reamplified a 2.0-kb product from the 3.2-kb product, by using the following internal primers: forward, 5′-CCCAAAGTCCCCACTATGTC-3′ (nucleotide 74531); and reverse, 5′-GCCAAGACG-

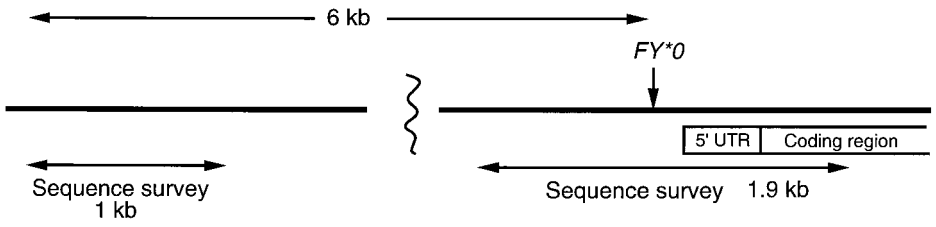

Figure 2 Duffy region, showing locations of the sequence surveys, with respect to the *FY*O* mutation

GGCACCACAAT-3′ (nucleotide 76566). All individuals were sequenced from nucleotide 74587 to nucleotide 76517. In the course of sequencing the Italian DNAs, we became aware that there was an excess of homozygotes. The problem was traced to a polymorphism at the target sequence of the forward primer for the 3.2-kb product (at nucleotide 74131): we had inadvertently been doing allele-specific PCR (AS-PCR). Genomic DNAs were screened for the mismatch, by sequencing, and the 2.0-kb product was amplified directly from genomic DNA, for all individuals carrying the mismatch. The mismatched allele was found in the Italian sample but not in any of the African samples.

When more than one site was heterozygous in one individual, haplotype phase was determined through a combination of approaches. For Beti individuals 16, 18, and 19, the 3.2-kb PCR product was cloned by use of the TOPO XL PCR cloning kit (Invitrogen), and two or more individual clones were sequenced. Haplotypes of the Mbuti Pygmies and the Luo were inferred to be the same as that of the Beti having the same genotype, since (*a*) both major haplotypes observed in the Beti were observed in homozygous state in the Mbuti Pygmies and (*b*) no new ambiguous genotypes were observed.

Sequences of Italian individuals 4, 5, 8, 10, 11, 12, 13, 14, and 16 were apparently homozygous when 3.2-kb PCR products were used as the template. These individuals were subsequently found to be heterozygous at nucleotide 74131 (see above) and were resequenced by use of truly diploid 2.0-kb PCR products. However, the sequence of the AS-PCR product provided haplotype information for these individuals. Italian individuals 1, 7, and 15 are truly homozygous. Haplotypes of Italian individuals 2, 3, 6, 9, and 17 were determined by cloning.

Analysis of the Region 5–6 kb 5′ to the FY*O *Mutation*

Two PCR steps were used to obtain clean amplification products for this region. The first round of PCR used forward primer 5′-ATTTGGGACTCTGGCTGTGG-3′ (nucleotide 69524) and reverse primer 5′-CCTCACCTGACCACACCCTTC-3′ (nucleotide 71414). A small amount of the first PCR product was used as template to amplify a 1.5-kb product, by use of forward primer 5′-CTGCTGATGTCTTTGCCACAC-3′ (nucleotide 69544) and reverse primer 5′-CGCAGGAAGCAATGAGATGG-3′ (nucleotide 70994). Because of a length variation in the product, we sequenced two noncontiguous regions of this PCR product—nucleotides 69583–69982 and nucleotides 70283–70882—to obtain 1 kb of sequence.

Haplotypes for each individual were inferred by the method of Clark (1990). For example, Beti individual 14, Hausa individual 4, and Hausa individual 17 (whose genotypes are identical) were assumed to have both one copy of the haplotype observed in homozygous state in seven individuals from these populations and a new haplotype present only in heterozygous state. Haplotype information for the Italian individuals, in whom three haplotypes were observed unambiguously, was inferred in the same way. Correct assignment of haplotypes in this region was not critical to any of the analyses performed on the Italian sample.

Data Analysis

Statistical analyses were performed by the program DNAsp, version 3.14 (Rozas and Rozas 1997).

Results

To characterize the signature of directional selection on patterns of variation linked to the *FY*O* mutation in sub-Saharan Africans, we surveyed sequence variation in a 1.9-kb region centered on the position of the mutation defining the *FY*O* allele and in a 1-kb region 5–6 kb 5′ to the *FY*O* mutation (fig. 2). These regions were surveyed in small random samples of individuals from each of five sub-Saharan African populations in which the *FY*O* allele is at or near fixation: the Mandinka from The Gambia, the Beti and the Hausa from Cameroon, the Mbuti Pygmies from the Central African Republic, and the Luo from Kenya. These regions also were surveyed in a sample of 17 Italians, who were expected to have only *FY*B* and *FY*A* alleles, thus representing a population not affected by positive selection on the *FY*O* allele. One orangutan sequence was determined

Table 1

DNA Sequence Variation in the Duffy Region

Source	Sequence Variation[b]			
	5–6 kb 5′ to *FY*O* Mutation	Duffy Gene Region[c]		
	666777	777777	77	777777777
	999000	444555	55	555555666
	568468	788011	33	689999135
	906524	179179	33	770159841
	647084	440206	78	027565024
Orangutan	TCACCC	CCCACC	CT	TTCCG–AAG
Beti:				
5a	.....	.....	..	C.....
5b	.....	.....	..	C.....
14a	.....	.....	..	C.....
14b	C...T.	.....	..	C.....
16b	.....	.....	..	C.....
18b	.....	.....	..	C...T....
19b	.....	.....	..	C.....
16a	.....	...G..	--	CC.....
18a	.....	...G..	--	CC.....
19a	.....	...G..	--	CC.....
Hausa:				
4a	.....	.....	..	C.....
7a	.....	.....	..	C..C.....
7b	.....	.....	..	C.....
15a	.....	.....	..	C.....
15b	.....	.....	..	C.C.....
17a	.....	.....	..	C.....
19a	.....	.....	..	C.....
19b	.....	.....	..	C.....
4b	C...T.	.....	..	C.....
17b	C...T.	.....	..	C.....
Mbuti Pygmy:[c]				
1029a	.....	.....	..	C.....
1029b	.....	.....	..	C.....
1030a	.....	.....	..	C.....
1032a	.....	.....	..	C.....
1030b	.....	...G..	--	CC.....
1031a	.....	...G..	--	CC.....
1031b	.....	...G..	--	CC.....
1032b	.....	...G..	--	CC.....
1036a	.....	...G..	--	CC.T.....
1036b	.....	...G..	--	CC.T.....
Luo:				
2Tb	.....	.....	..	C.....
3Ca	.....	.....	..	C.....
3Cb	.....	.....	..	C.....
3Da	.....	.....	..	C.....
4Ua	.....	.....	..	C.....
4Ub	.....	.....	..	C.....
2Ta	.....	...G..	--	CC.....
3Db	.....	...G..	--	CC.....
Mandinka:				
258a	.....	.....	..	C.....
258b	.....	.....	..	C.....
423a	.....	.....	..	C.....
423b	.....	.....	..	C.....
442a	.....	.....	..	C.....
442b	.....	.....	..	C.....

Table 1 Continued

Source	Sequence Variation[b]			
	5–6 kb 5′ to *FY*O* Mutation	Duffy Gene Region[c]		
	666777	777777	77	777777777
	999000	444555	55	555555666
	568468	788011	33	689999135
	906524	179179	33	770159841
	647084	440206	78	027565024
511a	.....	.....	..	C.....
511b	.....	..T...	..	C.....
684a	.....	.....	..	C.....
684b	..G...	.....	..	C.....
Italian:				
11a	.G....	.....	..	.....G.
1a	.....	.....	..	.....T.G.
1b	.....	.....	..	.....T.G.
2b	.....	.....	..	.....T.G.
5a	.....	.....	..	.....T.G.
6a	.....	.....	..	.....T.G.
10a	...T..	.....	..	.....T.G.
12a	.....	.....	..	.....T.G.
15a	.....	.....	..	.....T.G.
15b	.....	.....	..	.....T.G.
16a	.....	.....	..	.....T.G.
2a	.....	.....	..	..T....G.
6b	C..T..	.....	..	..T....G.
14b	.....	.....	..	..T....G.
3b	.....	.....	..	..T.....
7a	.....	.....	..	..T.....
7b	...T..	.....	..	..T.....
8a	...T..	.....	..	..T.....
4a	...T..	.....	..	..T.....A
13a	.....	.....	..	..T.....A
14a	...T..	.....	..	..T.....A
17a	.....	.....	..	..T.....A
4b	C...T.	-...T.	..	..T.....
5b	C...T.	-...T.	..	..T.....
8b	C...T.	-...T.	..	..T.....
9a	C...T.	-...T.	..	..T.....
10b	C...T.	-...T.	..	..T.....
11b	C...T.	-...T.	..	..T.....
12b	C...T.	-...T.	..	..T.....
9b	C...TT	-G..TT	..	.C.....
13b	C...TT	-G..TT	..	.C.....
16b	C...TT	-G..TT	..	.C.....
3a	.....	.....	..	.C....G..
17b	.....	.....	..	.C....G..

NOTE.—Nucleotide 75670 is the site of the *FY*O* mutation. Nucleotide 76342 is the site of the *FY*A* mutation.

[a] The letters "a" and "b" appended to an individual's identification denote the two chromosomes from that individual.

[b] Haplotypes for sites 69596-70844 were inferred, not determined experimentally (see the Subjects and Methods section). The singleton G at 67401 has arbitrarily been assigned to 11a rather than to 11b; if this region contains a homologue of mouse gene BL2 (see the Results section), site 68425 would be a silent site in an exon, and all other sites would be intronic.

[c] Positions 74794–75670 are noncoding; positions 75872–76180 are 5′ UTR; and positions 76342 and 76514 are replacement polymorphisms.

to allow both inference of ancestral states at each polymorphic site and an estimate of sequence divergence. The orangutan genotype is homozygous for *FY*B,* the ancestral allele (Chaudhuri et al. 1995). The data are shown in table 1.

Sequence Variation at the Duffy Gene Region

All African individuals sampled were *FY*O* homozygotes. In the combined African sample (48 chromosomes), there were a total of five single-nucleotide polymorphisms (SNPs) and one length variant. Two of the SNPs (at nucleotides 75012 and 75872), as well as the length variant, are in absolute linkage disequilibrium, resulting in two distinct haplotypes at intermediate frequency (.3/.7), with no evidence of recombination between them. Of the 48 African chromosomes surveyed, 44 can be accounted for by these two major haplotypes. This haplotype structure is surprising for an area affected by a selective sweep, since only low-frequency variants are to be expected (see the Discussion, below). The three remaining polymorphisms are two singletons and a doubleton, each occurring in a different population. Both major haplotypes at the Duffy locus were observed in the Beti, the Mbuti Pygmy, and the Luo, whereas the Hausa and the Mandinka had only the more common haplotype. Because only 10 chromosomes were surveyed per population, the absence of the minor haplotype in the Mandinka and the Hausa may be due to chance. In fact, we calculated that, if the minor haplotype occurred at a frequency <25%, we would have >5% probability of observing no copies of this haplotype in a sample of 10 chromosomes.

In contrast to the pattern seen at most other loci in humans (Przeworski et al., in press), which show greater variability in Africans than in Europeans, the Italian sample was more diverse than the African sample. In 34 chromosomes there were eight SNPs and two length variants arranged into eight haplotypes requiring at least two recombination events. In addition to the *FY*A*/*FY*B* amino acid polymorphism, there is an alanine-to-threonine polymorphism at nucleotide 76514, segregating within the *FY*B* allelic class. This variant has been observed before and has been estimated to occur at a frequency of ~16% in Europeans but to be absent in blacks from South Africa (Olsson et al. 1998). The 20 *FY*B* chromosomes segregate at five SNPs and one length variant that do not show variation in the *FY*O* and *FY*A* chromosomes. One length variant (nucleotide 75995) is unique to the *FY*A* class (14 chromosomes), but no additional polymorphic sites are observed. Thus, unlike other nuclear loci, the Duffy locus shows independent sets of variation in the African and Italian populations: only one polymorphism is shared between the samples, most likely because of recombination or gene conversion (see the Discussion, below).

Sequence Variation 5–6 kb 5′ to the FY*O *Mutation*

For all individuals surveyed at the Duffy gene region, we also surveyed 1 kb of DNA sequence starting 5 kb from the site of the *FY*O* mutation in the 5′ direction (fig. 2). There is no experimental evidence for functional genes in this region, although there is significant similarity to a mouse BL2 gene. If this similarity is indeed due to the presence of a functional gene, 122 nucleotides of the region surveyed would be in an exon and the remainder would be in an intron—that is, ⩾90% of the sites would be expected to be silent. (This exon is the 9th of 10 exons; the putative coding region begins at nucleotide 42543 and ends at nucleotide 71699). Furthermore, there are, between human and orangutan, no fixed amino acid differences in this exon. This implies that it is unlikely that this exon has been a target of directional selection in humans, but we cannot exclude the possibility that other portions of this putative gene have been exposed to positive natural selection.

Similar to what is observed in the region of the Duffy gene, sequence variation in this region is greater in the Italian sample, in which five SNPs are observed, three of which are at intermediate frequency (table 1). Of the three SNPs observed in the Africans, one is a singleton in the Mandinka. The other two SNPs are found together in a haplotype that is common in the Italian sample, suggesting that they became associated with *FY*O* alleles through a single recombination event with one of the non-*FY*O* alleles prior to their disappearance because of selection.

Levels of Polymorphism and Sequence Divergence

Estimates of polymorphism (π and θ_w) and of sequence divergence across both regions are shown in table 2, for population samples as well as for each allelic class. Under neutrality, π and θ_w are estimates of the quantity $4Ne\mu$, where Ne is the effective population size and μ is the neutral mutation rate. Estimates of μ, based on sequence divergence in humans vis-à-vis orangutan, are 1.04×10^{-9}/site/year and 1.14×10^{-9}/site/year for the Duffy gene region and the 5–6 kb 5′ region, respectively. These estimates are within the range that other studies have reported for this species pair (Bailey et al. 1991; Nachman et al. 1998), suggesting that the mutation rate in these regions is not unusual.

In both Africans and Italians, levels of polymorphism observed at the Duffy gene region and 5–6 kb away are very similar: at both regions, polymorphism levels are approximately two to three times higher in Italians than in Africans. This finding suggests that the signature of selection around the *FY*O* site extends over >6 kb. Note also that, in both regions, the African population samples that have the highest polymorphism are those for which there is evidence of recombination, as revealed by the presence of the same SNPs on both *FY*O* and *FY*B*

Table 2

Summary Statistics of Population Variation

		5–6 kb 5′ to *FY*O* Mutation[b]			Duffy Gene Region[c]			Combined Regions		
	N[a]	θ[d]	π[e]	D[f]	θ[d]	π[e]	D[f]	θ[d]	π[e]	D[f]
Beti	10	.71	.36	−1.40	.55	.59	.25	.60	.52	.53
Hausa	10	.71	.71	.02	0	0	0	.24	.24	.02
Mbuti Pygmies	10	0	0	0	.55	.74	1.23	.36	.48	1.23
Luo		0	0	0	.40	.44	.41	.26	.29	.41
Mandinka	10	.04	.02	−1.11	.18	.09	−1.11	.24	.14	−1.40
Total *FY*O*	48	.68	.39	−1.22	.58	.46	−.52	.62	.40	.97
Italian	34	1.22	1.43	.44	1.01	1.25	.67	1.08	1.31	.68
Total *FY*A*	14		…[g]		.16	.19	.32		…[g]	
Total *FY*B*	20		…[g]		1.02	1.23	.68		…[g]	
		Divergence (%)								
Human-orangutan		3.2			2.9			3.0		

[a] N = number of chromosomes.
[b] A total of 1,000 sites were surveyed.
[c] A total of 1,931 sites were surveyed.
[d] Watterson's (1975) estimate of $4N\mu$ (× 1,000).
[e] Nucleotide diversity (× 1,000).
[f] Tajima's (1989*b*) D statistic.
[g] Because we could not determine haplotypes across the entire region in the Italian sample, the 5′ variation was not classified as either *FY*A* or *FY*B*.

chromosomes. For example, the Beti have the highest polymorphism levels among the African samples, and they also show the two SNPs at nucleotides 69596 and 70628 in the 5′ region and the SNP at nucleotide 75872 in the Duffy gene region; these SNPs are also segregating in the Italian sample. Conversely, the Mandinka have the lowest polymorphism levels among the African samples and show no SNPs in common with the Italian sample. This suggests that recombination between *FY*O* and non-*FY*O* alleles, which must have occurred prior to the disappearance of non-*FY*O* alleles, has been more important than mutation has been, in the generation of the heterogeneity of *FY*O* chromosomes in these population samples.

Tests for a Departure from Neutral Equilibrium

Under neutrality, the amount of intraspecies polymorphism at a locus is proportional to the amount of interspecies divergence, since both are functions of the same underlying rate of neutral mutation (Kimura 1983). Directional selection, in contrast, is expected to reduce the level of neutral variation relative to the divergence. The HKA test of neutrality is based on this expectation (Hudson et al. 1987). It requires a neutral reference locus for which both polymorphism and divergence data have been obtained: under the null hypothesis, data for both loci should be consistent with the same estimates of effective population size and divergence time. We have used intron 44 of the *DMD* locus (M. W. Nachman and S. L. Crowell, personal communication) as the reference locus because divergence to orangutan had been determined for this region; several other polymorphism data sets are available with chimpanzee as an outgroup, but the very low divergence between chimpanzee and human provides little power in the HKA test (Kreitman 1991). By pooling across populations and sequenced regions, we found, in the African Duffy gene data, a significant departure from the neutral expectation (table 3). We interpret this result as reflecting a deficiency of polymorphism at the Duffy locus. When individual African population samples are tested, P values are $<.05$ for all samples except the Beti ($P = .11$). The Italian data, in contrast, are consistent with the null model, as might be expected, since selection at *FY*O* has not affected the Italian population.

Directional selection is also expected to generate a frequency spectrum skewed toward rare alleles. This is because a selective sweep is effectively a locus-specific bottleneck followed by population expansion, creating a "star" genealogy on which mutations will be disproportionately present as singletons (Slatkin and Hudson 1991). Tajima's D is a statistic based on the difference between θ_w, which is based on the number of segregating sites, and k, the average number of pairwise differences between sequences. Under neutrality and equilibrium, these two measures have equal expectations, and Tajima's D will be close to 0, but, after directional selection, the excess of singleton mutations will cause Tajima's D to be negative (Tajima 1989*a*, 1989*b*). In our analysis (table 2), D values are positive for the Italian

Table 3

Tests of Polymorphism and Divergence

	DMD[a]		DUFFY[b]						
	African	European	All Africans	Beti	Hausa	Mbuti Pygmy	Luo	Mandinka	Italians
Sample size	10	10	48	10	10	10	8	10	34
No. of segregating sites	14	10	8	5	2	3	2	2	13
P			.03	.11	.02	.03	.03	.02	.47

[a] The total number of sites is 3,000; the divergence is 78. Data are from M. W. Nachman and S. L. Crowell (personal communication).

[b] The total number of sites is 2,873 and consists of 1,931 sites at the Duffy locus plus 1,000 sites from the region 5′ to the Duffy locus minus 58 sites absent in orangutan; the divergence is 86.

sample and are negative for the total African sample, but only the Mandinka sample has negative *D* values for both regions surveyed. Although the magnitude of some of these *D* values is large, none of them is significantly different from 0. The power of Tajima's *D* to detect the effects of selection is weak when the sample sizes and the numbers of segregating sites are small (Simonsen et al. 1995).

As discussed above, a selective sweep is expected to result in an excess of singleton mutations. The Fu and Li (1993) *D* statistic specifically tests whether the number of singleton mutations is consistent with neutrality, given the total number of mutations. Of the African samples, only the Mandinka, with two singletons, have a *D* that approaches significance ($D = -1.808$; $P \sim .10$). None of the other samples shows a departure from the neutral expectation.

Discussion

In most empirical population-genetic studies, theoretical expectations are used to test a hypothesis about the data: failure to reject the null hypothesis is taken as lack of support for the alternative hypothesis. Here, however, prior independent evidence had led to the conclusion that directional selection had occurred at the Duffy locus. Therefore, in this study, we have contrasted our empirical data to the predictions of a particular model (i.e., selective sweep) of the effects that directional selection has on patterns of linked variation. Our survey of sequence variation around the *FY*O* mutation has shown that directional selection has brought about a significant reduction of polymorphism levels at the Duffy locus and that the signature of natural selection appears to extend ≥6 kb away from the selected site. Nevertheless, our results do not conform completely to theoretical expectations for the pattern of variation at a locus that has experienced a simple selective sweep. In fact, in addition to several low-frequency variants, we observe two major haplotypes at intermediate frequencies.

Prior evidence for directional selection on the *FY*O* allele comes from the unusual and well-documented pattern of geographic differentiation: the *FY*O* allele is essentially fixed in much of sub-Saharan Africa, although it is virtually absent in non-Africans. In addition to its striking geographic distribution, the *FY*O* allele is associated with complete resistance to vivax malaria. Although there is some uncertainty as to whether the agent that drove the *FY*O* allele to fixation in sub-Saharan Africa was indeed *P. vivax* or some other pathogen (see below), the role of selection is not in question. In addition, a predicted gene with 45% protein-sequence similarity to the mouse *BL2* gene (a member of the immunoglobulin superfamily) recently has been identified as being closely linked to the Duffy locus. No experimental studies have been done on either the mouse *BL2* gene or its human homologue; if functional, these genes may encode cell adhesion molecules. On the basis of the available data, it is hard to rule out the possibility that an advantageous mutation at this neighboring gene was the true target of selection and that the *FY*O* allele merely "hitch-hiked" to fixation. However, in light of all the evidence, it seems more likely that selection acted on the *FY*O* mutation itself. For the purposes of this discussion, we assume that this is the case, although definitive proof remains to be obtained.

Since vivax malaria is prevalent in much of Asia, and since there appear to be no deleterious effects of the *FY*O* mutation that would cause selection against it in all other populations, the fixation of this allele presumably occurred subsequent to the major human migrations out of Africa. This puts an upper limit of ~60,000–100,000 years (Watson et al. 1997; Jorde et al. 1998 and references therein) on the age of the selective event. Population-genetics theory predicts that such a recent (i.e., 0.3–0.5*N* generations) selective sweep should leave a marked signature in the genomic region surrounding the target of selection. Variation is expected to be reduced compared with that in other parts of the genome, and mutations that do occur are expected to be at low frequency, resulting in a negative Tajima's *D* (Simonsen et al. 1995). In agreement with these expectations, our survey of sequence variation at

the Duffy locus has shown that variation is significantly reduced in the Africans as a whole, over a region that extends 6 kb away from the mutation. Although African populations typically harbor ~50% more genetic diversity than is seen in Europeans, the Duffy region in a sample of 34 Italian chromosomes is two- to threefold more variable than that in a sample of 48 African chromosomes.

Other aspects of the data, however, are not consistent with a simple selective-sweep model. In fact, the Mandinka population sample is the only one that looks like a textbook case of a selective sweep: only two mutations were observed, each occurring in a single individual, producing a strongly negative Tajima's *D* and a significant reduction in variation according to the HKA test. In contrast, two population samples failed to show strong evidence of selection at the Duffy locus when they were analyzed separately: the Mbuti Pygmy sample has a strongly positive Tajima's *D*, whereas the Beti sample fails to reject the neutral expectation by the HKA test. These apparent discrepancies in the data correlate with the presence of two major haplotypes in the Duffy region. For example, in the Mbuti Pygmy sample these two haplotypes have approximately equal frequencies, leading to a relatively high proportion of intermediate-frequency variants (and, hence, to a positive Tajima's *D*). Conversely, only one of the two major haplotypes is observed in the Mandinka sample and in the Hausa sample, leading to the expected reduction of polymorphism levels as well as to the skewness in the frequency spectrum of variation.

The relationship of the *FY*O* haplotypes to each other and to their ancestral *FY*B* alleles is ambiguous. Possible explanations for the occurrence of two major *FY*O* haplotypes include mutation accumulation and drift, recombination, gene conversion, recurrence of the *FY*O* mutation, and/or population structure (e.g., see the study by Slatkin and Wiehe 1998). Any of these possibilities violates the simplifying assumptions of the selective-sweep model. In figure 3, we present three minimum-mutation networks, including all the haplotypes for the Duffy gene region only (1.9-kb sequence survey), from all the population samples. These networks represent three possible (but not exhaustive) scenarios for the origins of the two major haplotypes. In network I, the ancestral *FY*O* haplotype has undergone a recombination event, a nucleotide substitution, and an insertion/deletion event, generating the second major haplotype. The other two networks explain the presence of two major *FY*O* haplotypes as being the result of either a gene-conversion event (network II) or recurrent mutation (network III), both involving the site of the *FY*O* mutation. For the two latter scenarios, the absence of a donor chromosome may appear problematic. However, most of the variation linked to the *FY*B* allele has

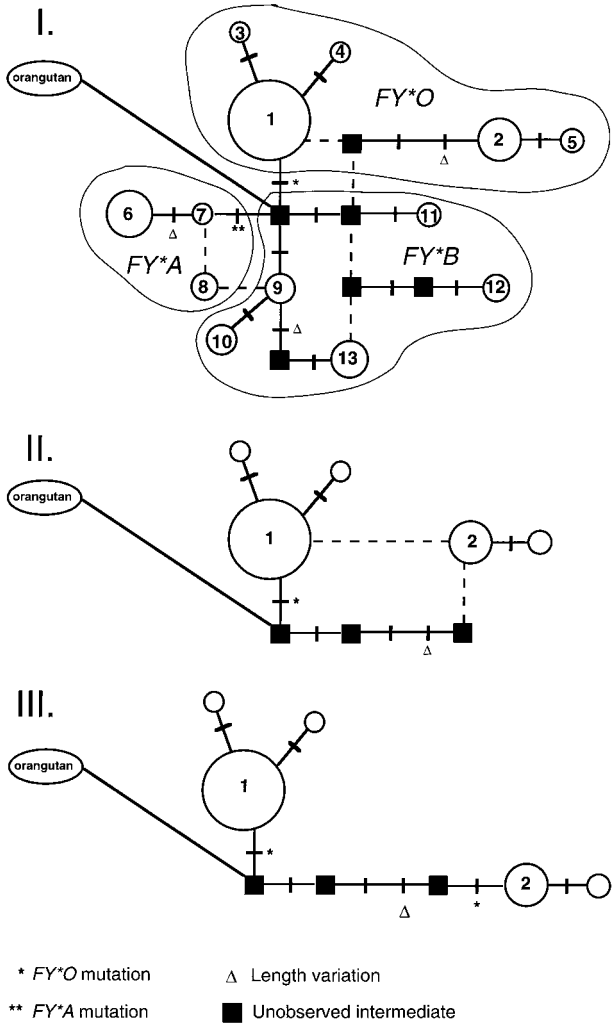

Figure 3 Possible minimum-mutation networks of the haplotypes observed in 1,931 bp of sequence around the *FY*O* site in Africans and Italians. Numbers in the circles denote the observed haplotypes. Representative individuals for each haplotype are from table 2: 1 = Beti individual 5; 2 = Beti individual 16a; 3 = Beti individual 18b; 4 = Mandinka individual 511b; 5 = Mbuti Pygmy individual 1036; 6 = Italian individual 1; 7 = Italian individual 11a; 8 = Italian individual 2a; 9 = Italian individual 7a; 10 = Italian individual 4a; 11 = Italian individual 3a; 12 = Italian individual 9b; 13 = Italian individual 4b. The sizes of the circles are proportional to the haplotype frequencies. Ticks on a lineage represent mutational events; dashed lines represent possible recombination events. Network I shows the entire genealogy for the Duffy gene region, whereas networks II and III show only the genealogy of the *FY*O* haplotypes. In network I, the *FY*O* mutation is assumed to have occurred once, and mutations at nucleotides 75082 and 75337 are assumed to have occurred on one lineage of the *FY*O* branch, along with a recombination event at nucleotide 75872 (shared with *FY*B*). In network II, mutations at nucleotides 75082 and 75337 are assumed to have occurred on an *FY*B* lineage that gained the *FY*O* allele through a gene-conversion event. Other possible networks involving multiple recombination events are not shown. In network III, a second *FY*O* mutation is assumed to have occurred on a different *FY*B* haplotype.

been lost from the African populations because of the selective fixation of the *FY*O* allele. Such lost variation may include the hypothetical donor chromosome; thus, it is not necessarily surprising that this chromosome was not observed in our samples.

The requirement for a donor chromosome in networks II and III implies that generation of the two major haplotypes by either gene conversion or recurrent mutation must have occurred prior to the selective fixation (as opposed to the beginning of the selection process) of the *FY*O* allele—that is, while *FY*B* chromosomes were still present. Even if network I is correct, three events are more likely to fall on one branch if both *FY*O* haplotypes were generated prior to selection and the intermediate haplotypes were lost by drift. Given that the *FY*O* mutation is recessive, the period of selection is likely to have been preceded by one of effective neutrality when the allele(s) was present only in heterozygotes. Substantial amounts of variation may have accumulated in this preselection phase. (This period of effective neutrality also violates the haploid or additive selective-sweep model, in which a new advantageous mutation is immediately subject to positive selection.)

Therefore, regardless of which of these scenarios is correct, any of the events that account for the presence of two major haplotypes almost certainly occurred prior to the fixation of the *FY*O* allele. In contrast, the three low-frequency variants (nucleotides 74890, 75915, and 75956) probably arose after the fixation. Therefore, the age of the selective episode (as opposed to the age of the *FY*O* mutation) is likely to be more accurately reflected by the low-frequency, rather than the intermediate-frequency, variants. If we assume, therefore, that only the three low-frequency variants have occurred since the time of selection, we can use those variants to obtain a rough estimate of the age of the selection on *FY*O*. We assume a star-shaped genealogy, as expected after a sweep: all the branches share a common ancestor at the time of fixation, and no coalescences have occurred since then. This star genealogy has 47 branches, one for each of the *FY*O* chromosomes surveyed (we count the two chromosomes of Mbuti Pygmy individual 1036 only once, since the homozygosity of the rare mutation at nucleotide 75915 in this individual suggests nonrandom mating). In the Duffy gene region (1,931 bp), there are three mutations falling on the 47 branches, and, hence, the average branch length is 3 mutations/47 branches = 0.0638 mutations/branch. We assume a mutation rate of 54 nucleotide substitutions/(2 × 14) million years divergence time between human and orangutan (Goodman et al. 1998)—that is, 1.93×10^{-6} mutations/year. Thus, the average branch length, in years, is 0.0638 mutations/branch/1.93×10^{-6} mutations/year = 33,075 years/branch. Under the assumption that the mutational process follows a Poisson distribution, the 95% confidence interval is 6,500–97,200 years.

A selective sweep occurring <97,200 years ago is consistent with hypotheses about human migrations out of Africa, migrations that are thought to have occurred ~100,000 years ago, although more-recent migrations also may have occurred (Watson et al. 1997). If the *FY*O* allele had been fixed in Africa >100,000 years ago, it would be present in significant frequencies in non-African populations today, unless it had experienced strong negative selection. There is some evidence that infection with vivax malaria may be advantageous in regions of high incidence of falciparum malaria, perhaps because it provides limited immunity to subsequent infection with *P. falciparum* (Williams et al. 1996). Furthermore, another study has shown that density-dependent regulation of parasite populations may result in inhibition of infection by *P. falciparum* simultaneous with infection by *P. vivax* (Bruce et al. 2000). However, these effects do not seem to be strong and are unlikely to explain the near absence of *FY*O* outside sub-Saharan Africa.

Whether a selective sweep occurring 33,000 years ago is consistent with the origins of malaria as a serious selective pressure in human evolution is less clear. Most hypotheses about the history of malaria in human populations have focused on falciparum malaria, which causes more mortality than does vivax malaria. The low diversity in malaria-resistance alleles such as β-thalassemia suggests that they are relatively young (Hill and Motulsky 1999), and it has been argued that falciparum malaria became an important human disease only after the development of agriculture, when human population densities increased (Livingstone 1984). This argument may not apply to vivax malaria, however, which can have an incubation period of 6–9 mo and which may relapse for as long as 3 years after exposure (Harinasauta and Bunnag 1998). Thus, vivax malaria is much less dependent on high host density and may have been an important cause of morbidity long before agriculture (Coluzzi 1996).

Although our estimates of the time of fixation of *FY*O* may be consistent with selection by vivax malaria, this explanation may still be questioned. On the basis of both the likely origin of *P. vivax* in Asia and the present-day distributions of vivax malaria and the *FY*O* allele, Livingstone (1984) argued that it was the prior fixation of the *FY*O* allele that fortuitously prevented the establishment of *P. vivax* in sub-Saharan Africa, rather than the opposite. This argument suggests a possible parallel between the *FY*O* allele and the CCR5-Δ32 deletion (Hill and Motulsky 1999). In the case of CCR5-Δ32, an unknown selective agent in the past is proposed to have caused an increase in frequency of the mutant allele, which happens, by chance, to con-

fer resistance to HIV-1 today (Stephens et al. 1998). The *FY*O* mutation could have a similar history—namely, that its frequency increased because of selection by a pathogen other than *P. vivax*. Receptors that are exposed on the surface of blood cells, such as the Duffy antigen and CCR5, are targets for exploitation by many pathogens (Pease and Murphy 1998) and may be subject to periodic episodes of selection by different agents.

The characterization of the signature of natural selection also may have important implications for the use of SNPs in disease-mapping studies. SNP-based approaches require the availability of a dense map of highly polymorphic markers, and, therefore, processes that affect the amount and distribution of sequence variation in the human genome are of great interest. Our study of the Duffy locus has shown that the number of polymorphic sites can be substantially reduced because of directional selection, even if the assumptions of the simple selective-sweep model are violated. This effect extends to the entire length (6 kb) of the surveyed region, suggesting that disease-mapping studies involving a region that experienced directional selection may be hampered by the low levels of variation. In addition, because the selective pressures differ across populations, levels and patterns of variation may show dramatic differences across populations, as a direct result of natural selection. In fact, at the Duffy locus, the non-African populations show different and more numerous SNPs than do the African populations, suggesting that SNPs identified by the screening of non-African samples would be of little use in the analysis of patients of African origin.

Most studies of human genetic variation have focused on deleterious alleles that are associated with disease. Our study of the Duffy locus has demonstrated the importance of another class of mutation, the targets of directional selection, and provides the first picture of the signature of that selection. Already it is clear that this signature may be complex—and that other loci will need to be characterized in order to have a full understanding of the process of directional selection. In addition, more-realistic models, notably those that incorporate recombination and population structure, need to be developed. These studies promise to provide important insights into the evolutionary history of our species.

Acknowledgments

We thank J. Donfack, G. Galluzzi, A. V. S. Hill, and J. Kidd for providing DNA samples; C. F. Aquadro, R. Harding, A. V. S. Hill, R. Hudson, M. Przeworski, and C.-I. Wu for discussion of the data and/or comments on the manuscript; M. W. Nachman for providing the *DMD* data prior to publication; and G. Ybazeta and A. Bartoszewicz for technical help. This work was supported by National Institutes of Health grant HG02098-02 to A.D.

Electronic-Database Information

The accession number and URL for data in this article are as follows:

GenBank, http://www.ncbi.nlm.nih.gov/Genbank/GenbankOverview.html (for bacterial-artificial-chromosome clone bk134P22 [accession number AL035403])

References

Aguade M (1998) Different forces drive the evolution of the *Acp26Aa* and *Acp26Ab* accessory gland genes in the Drosophila melanogaster species complex. Genetics 150: 1079–1089

Bailey WJ, Fitch DH, Tagle DA, Czelusniak J, Slightom JL, Goodman M (1991) Molecular evolution of the $\psi\eta$-globin gene locus: gibbon phylogeny and the hominoid slowdown. Mol Biol Evol 8:155–184

Bruce MC, Donnelly CA, Alpers MP, Galinski MR, Barnwell JW, Walliker D, Day KP (2000) Cross-species interactions between malaria parasites in humans. Science 287:845–848

Cavalli-Sforza LL, Menozzi P, Piazza A (1994) The history and geography of human genes. Princeton University Press, Princeton, NJ

Chaudhuri A, Polyakova J, Zbrzezna V, Pogo AO (1995) The coding sequence of Duffy blood group gene in humans and simians: restriction fragment length polymorphism, antibody and malarial parasite specificities, and expression in non-erythroid tissues in Duffy-negative individuals. Blood 85:615–621

Clark AG (1990) Inference of haplotypes from PCR-amplified samples of diploid populations. Mol Biol Evol 7:111–122

Coluzzi M (1996) Interazione evolutive Uomo-Plasmodio-Anophele. In: XXIII Seminario sulla "Evoluzione biologica e i grandi problemi della biologia." Accademia Nazionale dei Lincei, Rome, pp 263–285

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

Goodman M, Porter CA, Czelusniak J, Page SL, Schneider H, Shoshani J, Gunnell G, et al (1998) Toward a phylogenetic classification of primates based on DNA evidence complemented by fossil evidence. Mol Phylogenet Evol 9:585–598

Hadley TJ, Peiper SC (1997) From malaria to chemokine receptor: the emerging physiologic role of the Duffy blood group antigen. Blood 89:3077–3091

Harinasauta T, Bunnag D (1988) The clinical features of malaria. In: Wernsdorfer W, McGregor I (eds) Malaria: principles and practice of malariology. Churchill Livingstone, London, pp 709–734

Hill AVS, Motulsky AG (1999) Genetic variation and human disease: the role of natural selection. In: Stearns SC (ed) Evolution in health and disease. Oxford University Press, Oxford, pp 50–61

Hudson RR (1990) Gene genealogies and the coalescent process. Oxf Surv Evol Biol 7:1–44

Hudson RR, Kreitman M, Aguade M (1987) A test of neutral

molecular evolution based on nucleotide data. Genetics 116: 153–159

Jorde LB, Bamshad M, Rogers AR (1998) Using mitochondrial and nuclear DNA markers to reconstruct human evolution. BioEssays 20:126–136

Karn RC, Nachman MW (1999) Reduced nucleotide variability at a androgen-binding protein locus (*Abpa*) in house mice: evidence for positive natural selection. Mol Biol Evol 16:1192–1197

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

Kreitman M (1991) Detecting selection at the level of DNA. In: Selander RK, Clark AG, Whittam TS (eds) Evolution at the molecular level. Sinauer Associates, Sunderland, MA, pp 204–221

Livingstone FB (1984) The Duffy blood groups, vivax malaria, and malaria selection in humans: a review. Hum Biol 56: 413–425

Maynard Smith J, Haigh J (1974) The hitch-hiking effect of a favourable gene. Genet Res 23:23–35

Metz EC, Palumbi SR (1996) Positive selection and sequence rearrangements generate extensive polymorphism in the gamete recognition protein bindin. Mol Biol Evol 13:397–406

Metz EC, Robles-Sikisaka R, Vacquier VD (1998) Nonsynonymous substitution in abalone sperm fertilization genes exceeds substitution in introns and mitochondrial DNA. Proc Natl Acad Sci USA 95:10676–10681

Miller LH, Mason SJ, Clyde DF, McGuiness MH (1976) The resistance factor to *Plasmodium vivax* in Blacks: the Duffy-blood-group genotype, FyFy. N Engl J Med 295:302–304

Nachman MW, Bauer VL, Crowell SL, Aquadro CF (1998) DNA variability and recombination rates at X-linked loci in humans. Genetics 150:1133–1141

Olsson ML, Smythe JS, Hansson C, Poole J, Mallinson G, Jones J, Avent ND, et al (1998) The Fy^x phenotype is associated with a missense mutation in the Fy^b allele predicting Arg89Cys in the Duffy glycoprotein. Br J Haematol 103: 1184–1191

Pease JE, Murphy PM (1998) Microbial corruption of the chemokine system: an expanding paradigm. Semin Immunol 10: 169–178

Przeworski M, Hudson R, Di Rienzo A. Adjusting the focus on human variation. Trends Genet (in press)

Rozas J, Rozas R (1997) DnaSP version 2.0: a novel software package for extensive molecular population genetics analysis. Comput Appl Biosci 13:307–311

Simonsen KL, Churchill GA, Aquadro CF (1995) Properties of statistical tests of neutrality for DNA polymorphism data. Genetics 141:413–429

Slatkin M, Hudson RR (1991) Pairwise comparisons of mitochondrial DNA sequences in stable and exponentially growing populations. Genetics 129:555–562

Slatkin M, Wiehe T (1998) Genetic hitch-hiking in a subdivided population. Genet Res 71:155–160

Stephens JC, Reich DE, Goldstein DB, Shin HD, Smith MW, Carrington M, Winkler C, et al (1998) Dating the origin of the CCR5-Δ32 AIDS-resistance allele by the coalescence of haplotypes. Am J Hum Genet 62:1507–1515

Tajima F (1989*a*) The effect of change in population size on DNA polymorphism. Genetics 123:597–602

——— (1989*b*) Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. Genetics 123: 585–595

Tournamille C, Colin Y, Cartron JP, Le Van Kim C (1995) Disruption of a GATA motif in the Duffy gene promoter abolishes erythroid gene expression in Duffy-negative individuals. Nat Genet 10:224–228

Tsaur S-C, Ting C-T, Wu C-I (1998) Positive selection driving the evolution of a gene of male reproduction, *Acp26Aa* of *Drosophila:* divergence versus polymorphism. Mol Biol Evol 15:1040–1046

Wang R-L, Stec A, Hey J, Lukens L, Doebley J (1999) The limits of selection during maize domestication. Nature 398: 236–239

Watson E, Forster P, Richards M, Bandelt H-J (1997) Mitochondrial footprints of human expansions in Africa. Am J Hum Genet 61:691–704

Watterson GA (1975) On the number of segregating sites in genetical models without recombination. Theor Popul Biol 7:256–276

Williams T, Maitland K, Bennett S, Ganczakowski M, Peto TEA, Newbold CI, Bowden DK, et al (1996) High incidence of malaria in α-thalassaemic children. Nature 383:522–525