Am. J. Hum. Genet. 70:369–383, 2002

Complex Signatures of Natural Selection at the Duffy Blood Group Locus

Martha T. Hamblin, Emma E. Thompson, and Anna Di Rienzo

Department of Human Genetics, University of Chicago

The Duffy blood group locus (*FY*) has long been considered a likely target of natural selection, because of the extreme pattern of geographic differentiation of its three major alleles (*FY*B, FY*A,* and *FY*O*). In the present study, we resequenced the *FY* region in samples of Hausa from Cameroon (fixed for *FY*O*), Han Chinese (fixed for *FY*A*), Italians, and Pakistanis. Our goals were to characterize the signature of directional selection on *FY*O* in sub-Saharan Africa and to understand the extent to which natural selection has also played a role in the extreme geographic differentiation of the other derived allele at this locus, *FY*A*. The data from the *FY* region are compared with the patterns of variation observed at 10 unlinked, putatively neutral loci from the same populations, as well as with theoretical expectations from the neutral-equilibrium model. The *FY* region in the Hausa shows evidence of directional selection in two independent properties of the data (i.e., level of sequence variation and frequency spectrum), observations that are consistent with the *FY*O* mutation being the target. The Italian and Chinese *FY* data show patterns of variation that are very unusual, particularly with regard to frequency spectrum and linkage disequilibrium, but do not fit the predictions of any simple model of selection. These patterns may represent a more complex and previously unrecognized signature of positive selection.

Introduction

Under evolutionary neutrality, variation in allele frequencies across subpopulations is determined simply by genetic drift. Because drift is determined entirely by the demographic properties of the populations, all loci in the genome have the same expected degree of differentiation, which can be summarized by the F_{ST} statistic. The action of natural selection, however, may increase the variance of F_{ST} for a group of loci: balancing selection or specieswide directional selection at some loci may generate a more uniform distribution of allele frequencies relative to that at neutral loci, whereas local adaptation may increase the level of differentiation at others (Cavalli-Sforza 1966; Lewontin and Krakauer 1973; Bowcock et al. 1991; Cavalli-Sforza et al. 1994). Thus, if allele-frequency data are available for a set of putatively neutral loci, the distribution of their F_{ST} values can be compared with the F_{ST} values of specific variants hypothesized to have evolved under positive natural selection (Karl and Avise 1992; Berry and Kreitman 1993; Taylor et al. 1995). When the Duffy blood group locus (*FY* [MIM 110700]) is compared with putatively neutral loci, the degree of geographic differentiation of its three major alleles (*FY*B, FY*A,* and *FY*O*) is clearly among the highest observed in humans. For this reason, *FY* has been considered a likely target of population-specific selection pressures.

The *FY*O* allele, which is fixed in sub-Saharan Africa but is essentially absent elsewhere, is also thought to have been the target of positive selection because it confers resistance to malaria due to *Plasmodium vivax* (Livingstone 1984; Tournamille et al. 1995; Hadley and Peiper 1997). The *FY*O* allele differs from the ancestral *FY*B* allele by a single, noncoding base change that eliminates transcription of the *FY* mRNA in erythroid cells (Tournamille et al. 1995). Recently, we studied DNA sequence variation underlying the *FY*O* serotype in several small population samples from sub-Saharan Africa and found a reduction of variation consistent with the predictions of models of directional selection (Hamblin and Di Rienzo 2000). Some aspects of the data, however, were inconsistent with simple selection models. More specifically, the *FY*O* allele was found on two different haplotypes that occurred at intermediate frequencies in the entire sample and whose frequencies appeared to vary from population to population.

The *FY*A* allele, which differs from the *FY*B* allele by a single amino acid, also has a very unusual level of geographic differentiation, namely, its near-fixation frequency in eastern Asia and the Pacific (Cavalli-Sforza et al. 1994). Thus, the ancestral *FY*B* allele disappeared from much of the human population, becoming restricted to western Asia, Europe, and the Americas. Fixation of an allele through random genetic drift is a slow process (Kimura 1983). Because an allele at high frequency is expected to be old and, thus, to occur on a

Received October 4, 2001; accepted for publication November 8, 2001; electronically published December 20, 2001.

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

0002-9297/2002/7002-0011$15.00

variable genetic background, it was surprising to find little variation in a sample of *FY*A* chromosomes (Hamblin and Di Rienzo 2000). The low level of variation in such a common allele, coupled with its unusually high F_{ST} value, prompted the question of whether a previously unrecognized selective pressure on the *FY*A* allele (or one tightly linked to it) underlies the observed patterns. Unlike the *FY*O* allele, the *FY*A* allele has not been associated with any phenotype.

Hence, both the *FY*O* and *FY*A* alleles are excellent candidates for use in the exploration of the effects of positive natural selection on patterns of linked neutral variation. Detection of the signature of natural selection requires one to distinguish between the effects of natural selection and those of population history. This is a particularly difficult challenge in humans for two main reasons. First, the low levels of sequence variation provide little power to detect a significant reduction of polymorphism levels, which would be expected under directional selection. Second, because human demographic history appears to have been complex (implying that populations are not at equilibrium), the theoretical expectations of an equilibrium model are not a valid null hypothesis for testing neutrality. To overcome the latter problem, one can compare the pattern of variation at positively selected loci to empirical—rather than theoretical—expectations derived from the analysis of patterns of variation at putatively neutral loci. Such empirical comparisons allow the identification of genomic regions with exceptional patterns of variation, regardless of whether the correct demographic model is known.

We present a survey of variation in the *FY*-gene region in population samples from sub-Saharan Africa (Hausa) and eastern Asia (Han Chinese), which are essentially fixed for the *FY*O* and *FY*A* alleles, respectively. Variation in these samples is compared with that in samples from Europe (Italians) and southern Asia (Pakistani) in which two or three of the *FY* alleles are segregating (i.e., in which a fixation has not occurred). The present survey was done in parallel with a companion study that used a large subset of exactly the same population samples (Frisse et al. 2001). The companion study focused on anonymous, noncoding genomic regions (referred to as the "locus pairs") that may provide a reasonable description of levels and patterns of neutral variation in these populations. The data from the *FY* region are compared with the patterns of variation observed at these putatively neutral loci, which alleviates the need for a theoretical description of the underlying population history. The results show a distinct signature of natural selection in the African sample and are consistent with an appreciable, but more complex, signature in the non-African samples.

Subjects and Methods

Population Samples

Sequence variation was surveyed in DNA samples from 16 Italians, 14 Pakistanis, 16 Hausa, and 16 Han Chinese. The 16 Italians and 5 of the Hausa are included in a report published elsewhere (Hamblin and Di Rienzo 2000). The present study was approved by the institutional review board of the University of Chicago.

PCR and Sequence Determination

PCR primers and sequencing primers were based on sequence from GenBank accession number AL035403. A list of primers used in this project is available from the authors. PCR products were prepared for sequencing by treatment with exonuclease I and shrimp alkaline phosphatase (United States Biochemicals). Dye-terminator sequencing was performed with ABI Big-Dye reagents, and products were analyzed on an ABI 377 or 3700 automated sequencer (Applied Biosystems). Sequence was determined on both strands. Chromatograms were imported into Sequencher 3.1.1 (Gene Codes) for assembly of contigs and identification of polymorphic sites. Contigs were initially assembled separately for each individual and were visually inspected for identification of heterozygous sites. Individual contigs were then compared within and across populations, for scoring of polymorphisms.

Data Analysis

Summary statistics of DNA sequence variation, divergence, and F_{ST} were calculated using DnaSP (Rozas and Rozas 1999). $|D'|$ and r^2 were calculated from diploid data, according to the maximum-likelihood method (Hill 1974). Only sites for which the rare allele was present on at least three chromosomes were included in the analysis.

To obtain a reference "locus" for the Hudson-Kreitman-Aguadé (HKA) test (Hudson et al. 1987), we used a pooled data set from nine unlinked, noncoding regions for which divergence data for an orangutan were available. (Although the HKA test assumes no recombination within loci, this assumption is conservative.) These regions, which were chosen according to a set of criteria intended to eliminate regions that may have been subject to natural selection, were surveyed in the same population samples of Hausa, Italians, and Chinese (Frisse et al. 2001). (HKA tests were not applied to the Pakistani data, because no suitable reference data set was available.)

Estimation of ρ

The population recombination parameter ρ was estimated from diploid data, as described by Frisse et al. (2001) (program available at the Hudson Laboratory

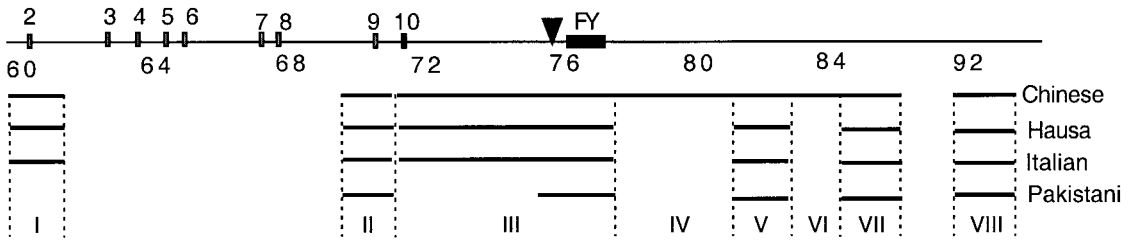

Figure 1 *FY*-gene region. Numbers (× 1,000) below the line refer to the nucleotide position from GenBank accession number AL035403. The shaded vertical bars (numbered above the line) indicate the positions of exons in a putative BL2-like gene 5′ to the *FY* locus. The inverted triangle indicates the position of the *FY*O* mutation 5′ to the *FY* coding region (*blackened horizontal bar*). Thick lines below the numbered line indicate the regions surveyed for sequence variation in each population. Roman numerals indicate how the sequence data were partitioned into segments for the purpose of data analysis. This partition was arbitrary and simply reflects the boundaries of the segments surveyed in different population samples.

Web site). The gene-conversion parameters used were as follows: a mean tract length of 500 and gene conversion to crossing-over rate (*f*) of 8. Polymorphisms for which the rare allele was present on at least three chromosomes (i.e., a frequency >9% in a sample of 32 chromosomes) were included in the analysis.

Haplotype Estimation

The haplotype phase of the sequence data for the Italian sample from positions 74587–76517 was determined by a combination of allele-specific (AS) PCR and cloning (Hamblin and Di Rienzo 2000). In the present study, the same polymorphism at 74131 was used in AS-PCR to determine the phase of alleles at positions 71784–73841. Thus, in individuals who were polymorphic at 74131 (8 of 16 Italians), phase was experimentally determined across the entire region of 71784–76517.

For the rest of the diploid data, the program PHASE (Stephens et al. 2001) was used to estimate haplotypes. This program assigns a pair of haplotypes for each diploid entry and provides probability scores for each unknown position. Singletons were eliminated from the analysis because they cannot be assigned with any confidence. Experimentally determined phase information for the Italian sample was incorporated into the analysis. Because the haplotypes estimated for the *FY* region were 23 kb, recombination is likely to have occurred on some chromosomes. Therefore, we also estimated the haplotypes on shorter, overlapping subsets of sites, which, in a few cases, resulted in improved probability scores for different configurations of alleles. In these cases, the larger haplotypes were constructed from an assembly of the haplotypes assigned to shorter regions with overlapping sites. The proportion of unknown sites for which the phase was assigned with >90% probability was 95% for the Chinese, 85% for the Hausa, 84% for the Italians, and 83% for the Pakistanis.

Results

A diagram of the *FY* region, indicating the regions surveyed in each population sample, is presented in figure 1. The regions cover ~7.5–18 kb (mean 12 kb) and span a distance of ~34 kb that is roughly centered on the *FY* locus. The vast majority of this sequence is noncoding; because we are interested in detecting the effects of directional selection (which is expected to affect all sites linked to an advantageous mutation, regardless of their function), we have not divided the sites into functional categories. Details of the regions surveyed and summary statistics of sequence variation are shown in table 1.

All 32 chromosomes in the Hausa sample bear *FY*O* alleles, and all 32 chromosomes in the Chinese sample bear *FY*A* alleles. Because the individual DNA samples were randomly selected, the results reflect the near-fixation frequencies of these alleles in these populations. The Italian sample consists of 18 *FY*B* alleles and 14 *FY*A* alleles, and the Pakistani sample consists of 12 *FY*B* alleles, 12 *FY*A* alleles, and 4 *FY*O* alleles.

Our interest in patterns of sequence variation at the *FY* locus arises from the very high F_{ST} values observed at this locus. If positive selection on the *FY*A* and *FY*O* mutations is responsible for these extreme values, the value of F_{ST} should diminish with distance from these sites. Figure 2 (*top*) shows the profile of F_{ST} between Italians and Chinese as a function of distance from the *FY*A* mutation, a comparison that is not influenced by the presence of the *FY*O* allele. Figure 2 (*bottom*) shows F_{ST} between Italians and Hausa as a function of distance from the *FY*O* mutation. Because there is no population that lacks both *FY*A* and *FY*O* alleles, the confounding effects of *FY*A* in the Italian sample cannot be avoided. Overall, F_{ST} values in this region are clearly higher than those observed at putatively neutral loci in the same population samples. In both comparisons, the highest F_{ST} values are observed closer to the

Table 1

Summary Statistics of Population Variation

Overall and Population Characteristics[a]	Bases 59366–60682	Bases 69583–70882[b]	Bases 71676–86020[c]					Bases 91843–93282
	I	II	III (FY)	IV	V	VI	VII	VIII
Length (nt)	1,311	1,000	5,547	3,912	1,214	1,942	1,727	1,440
Divergence	18	32	154	105	30	45	37	54
Chinese:								
S	1	4	5	9	0	6	4	4
π ($\times 10^{-3}$)	.13	.69	.20	.82	0	1.55	1.18	1.28
D	−.45	−.77	−.33	1.33	−	2.84[d]	2.64[d]	2.17[d]
No. of haplotypes	2	4	6	7	1	3	2	3
P (HKA)	.55	.81	.07	.66	.11	.74	.99	.55
P (*H*)	.67	.76	.24	.32	…	.23	.30	.25
Hausa:								
S	1	2	6	…	2	…	7	6
π ($\times 10^{-3}$)	.34	.54	.27		.50		.95	.90
D	1.20	.19	.05		.86		−.15	−.37
No. of haplotypes	2	2	6		3		3	7
P (HKA)	.32	.25	.02		.28		.83	.46
P (*H*)	.92	.94	.05		.006		.004	.59
Italian:								
S	3	5	16	…	10	…	4	6
π ($\times 10^{-3}$)	.23	1.46	.78		.82		1.12	1.06
D	−1.38	.44	.37		−2.14[d]		2.39[d]	.06
No. of haplotypes	4	6	17		4		3	6
P (HKA)	.55	.55	.79		.01		.99	.98
P (*H*)	.46	.94	.37		.00		.29	.82
Pakistani:								
S	…	4	10	…	1	…	7	8
π ($\times 10^{-3}$)		.66	1.34		.16		1.46	1.50
D		−1.07	.40		−.36		1.22	.15
No. of haplotypes		5	12		2		5	9
P (*H*)		.72	.63		.72		.43	.85

[a] Divergence is the number of differences between two randomly chosen alleles of human and orangutan. The length used to calculate divergence is the same as the length surveyed for polymorphism, except for regions III (5,188 nt), IV (3,861 nt), VI (1,877 nt), and VIII (1,410 nt). *S* is the number of segregating sites; π is nucleotide diversity; *D* is from Tajima (1989); *P* (HKA) is the probability of the χ^2 statistic of the HKA test using the pooled locus-pairs data as the reference locus; *P* (*H*) is the probability of the *H* statistic of Fay and Wu's test (2000) without recombination.

[b] Region II was sequenced in bases 69583–69982 and 70283–70882, because of length variation.

[c] The Pakistani sample was surveyed only from bases 75374–77525 (length 2,152 nt) in region III.

[d] $P < .05$.

putative selected mutation, although there is also a very high F_{ST} ~15 kb 3′ to the *FY*A* mutation.

The high degree of geographic structure at the *FY* locus is also reflected in the tree based on inferred haplotypes for region III (see Methods section and fig. 3). The Pakistanis were not included because not all the sites had been surveyed. To characterize the geographic structure of variation linked to the *FY*O* and *FY*A* sites, these two sites were omitted from the sequence haplotypes in this analysis. Unlike the trees observed at the vast majority of nuclear loci, there are no cosmopolitan haplotypes. In addition, the Hausa haplotypes cluster at the base of the tree. Likewise, the Chinese tend to cluster together. If the *FY*O* and the *FY*A* sites are included, all the Hausa fall into a single distinct clade, and the Chinese are restricted to a subset of the Italian genealogy.

The Signature of Natural Selection in the African Sample

In a report published elsewhere (Hamblin and Di Rienzo 2000), we showed that polymorphism levels around the *FY*O* mutation are lower than would be expected on the basis of a neutral-equilibrium model in sub-Saharan Africans. This was done by comparing the sequence variation at the *FY* locus with that observed at intron 44 of the *DMD* locus. Here, we attempt to characterize this signature for multiple aspects of the data and over larger distances. To this purpose, we surveyed a larger region around the *FY*O* mutation in a single, larger population sample and compared the results with those from a multilocus data set of noncoding sequence variation (the "locus pairs") (Frisse et al. 2001). That noncoding survey showed that the Hausa sample fits the equi-

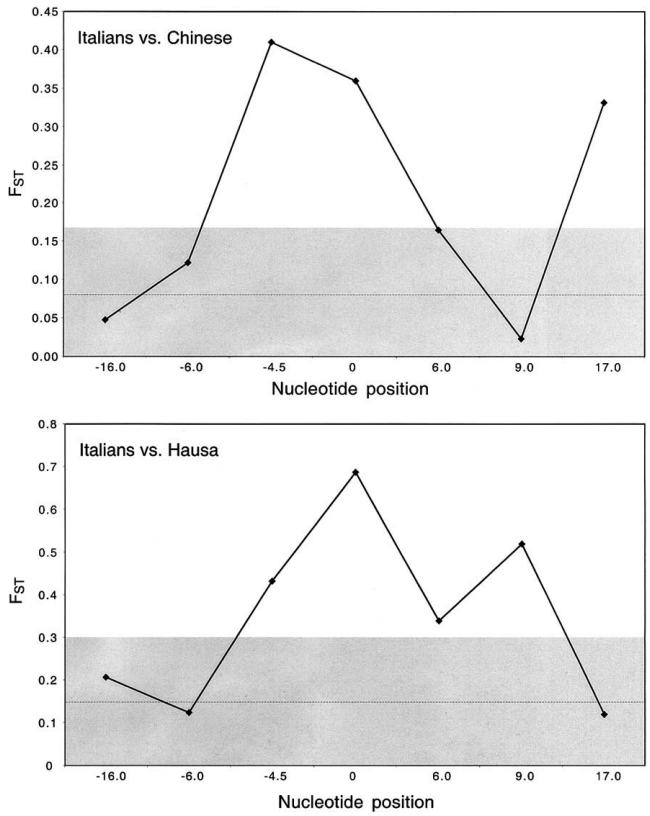

Figure 2 Comparison of F_{ST} in the *FY* region and in the 10 locus-pairs regions. Blackened diamonds represent F_{ST} at various distances from the *FY* gene. Position 0 represents the location of the *FY*A* mutation in the top panel and the *FY*O* mutation in the bottom panel. Regions II and III were pooled and then divided into three regions of ~2.5 kb, the average size of the locus-pairs regions. The shaded area represents the range of values of F_{ST} observed, between these populations, in the locus-pairs regions. The dashed line represents the average F_{ST} in the locus-pairs regions.

librium model for all aspects of the data. This implies that a comparison of the data at the *FY* region with the theoretical expectations from the equilibrium model is an appropriate test of natural selection for this population.

Simple models of directional selection predict that a gene genealogy will be shallower after the fixation of an advantageous mutation, which will result in an apparent local reduction of the effective population size (N_e) (Maynard Smith and Haigh 1974). Under the equilibrium model, N_e can be estimated from two parameters, each based on largely independent aspects of the data: the population mutation parameter θ ($\equiv 4N_e\mu$), which is based on polymorphism levels, and the population crossing-over parameter ρ ($\equiv 4N_e r$), which is based on linkage disequilibrium (LD). Estimates of μ, the mutation rate per site per generation, can be obtained from interspecific divergence data, and those of the crossing-over rate, r, between adjacent sites per generation, can be obtained from empirical measurements of crossing-over and physical distance. These estimates allow N_e to be calculated from each parameter estimate. In table 2, we compare estimates of N_e based on the two parameters in the *FY* region with average estimates based on the 10 locus pairs. The estimates of the population crossing-over rate are obtained using the maximum-likelihood values of gene-conversion parameters obtained for the Hausa (conversion to crossing-over rate [f] and the mean tract length; (Frisse et al. 2001; Hudson 2001) (see Methods section). In the Hausa sample, estimates of N_e based on either parameter are smaller in the *FY* region than in locus pairs. Interestingly, differences in N_e based on levels of polymorphism are less marked (as much as 1.7-fold) than those based on LD (4–8.5-fold), suggesting that the latter is more sensitive to directional selec-

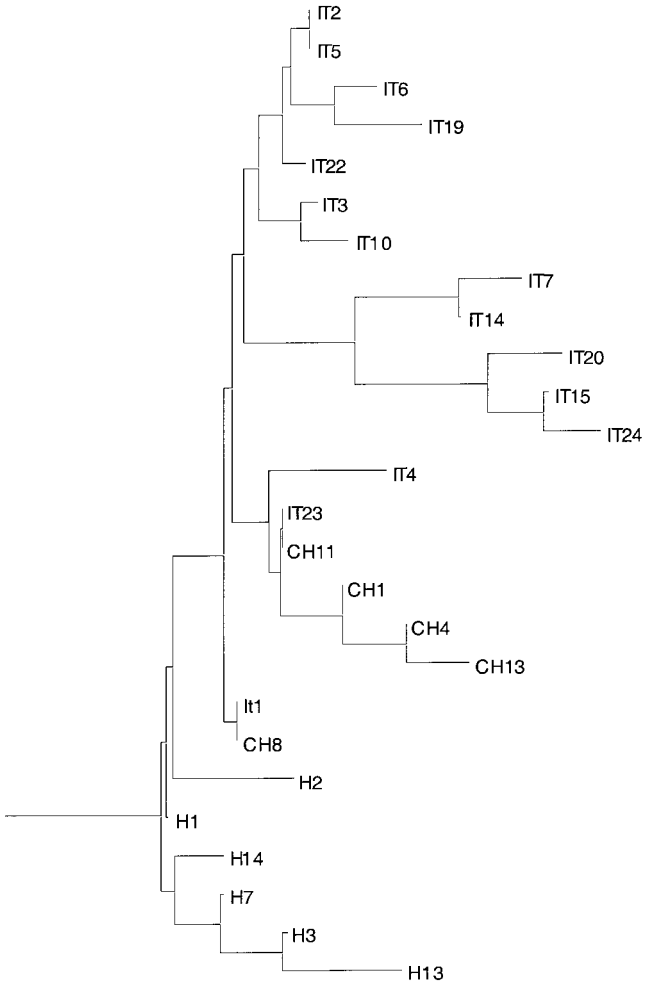

Figure 3 Neighbor-joining tree of region III, for the Hausa, Chinese, and Italians. Haplotypes inferred by PHASE were used in the analysis and are represented only once, even if they occurred multiple times in a sample. The *FY*O* and *FY*A* sites were omitted from the sequence of the inferred haplotypes.

Table 2

Estimates of Effective Population Size

	Estimates Based on					
	Recombination		Mutation			
			π		S	
Population and Test[a]	$\hat{\rho}$ ($\times 10^{-3}$)	N_e[b]	θ ($\times 10^{-3}$)[c]	N_e[d]	θ ($\times 10^{-3}$)[c]	N_e[d]
Chinese:						
LP	.2	4,457	.7	7,353	.8	8,328
FY	.7	1,259	.6	8,654	.4	5,907
Hausa:						
LP	.8	16,279	1.1	11,555	1.2	13,152
FY	2.5	4,496	.6	7,692	.6	8,104
Italians:						
LP	.2	3,682	1.0	10,504	.8	8,766
FY	4.1	7,374	1.0	13,462	1.0	13,599
Pakistani:						
LP	…	…	…	…	…	…
FY	3.5	6,295	1.1	15,385	1.0	14,011

[a] LP indicates locus pairs.

[b] Estimates for locus pairs are based on the average recombination rate of 1.29 cM/Mb; estimates for *FY* are based on a recombination rate of 1.39 cM/Mb (Nachman 2001).

[c] Estimates of θ for *FY* were based on the 7,533 bases sequenced in all four samples.

[d] N_e estimates are $\theta/4\mu$, where μ is based on divergence from orangutan: 2.3×10^{-8}/bp per generation for the locus pairs and 1.8×10^{-8}/bp per generation for *FY*, under the assumption of a divergence time of 14 million years and a generation time of 20 years.

tion. Estimates of N_e based on ρ for the Italian sample are higher in the *FY* region than in the locus pairs. Thus, the crossing-over rate in this region is not lower than that at the locus pairs and cannot account for the lower estimates of ρ in the Hausa.

It is possible to use the HKA test (see Methods section) (Hudson et al. 1987) to test whether the estimate of N_e based on polymorphism levels is significantly low. Levels of polymorphism and interspecific divergence at *FY* were compared with those in the pooled locus-pairs data. The results of these tests, by region, are shown in table 1. The Hausa have a deficiency of variation ($P = .02$) in region III, which spans the *FY*O* mutation. It should be pointed out that the test is based on the assumption of no recombination. This assumption is clearly violated, because the 10 locus pairs are unlinked, and recombination occurs within pairs (Frisse et al. 2001); as a consequence, the *P* values would be smaller if recombination were properly taken into account. HKA tests of the *FY* region III versus each individual locus pair (table 3) show that no *P* value is >.06 and that three are <.005. Thus, there is a clear reduction of variation in the vicinity of the *FY*O* mutation. Although the flanking regions do not depart when tested separately, a reduction of variation is also observed when regions I–III and V, which span 22 kb, are pooled ($P = .02$). At 10 kb 3′ to the *FY*O* mutation (region VII), however, there is no evidence of reduced variation, indicating that the signature of selection has dissipated <10 kb from the target of selection.

The frequency spectrum of mutations that arise after an episode of directional selection is expected to be skewed toward rare alleles, since the recovery from a selective episode is analogous to population growth after a bottleneck (Braverman et al. 1995). The *D* statistic (Tajima 1989), frequently used to assess such a skew, has an expectation near zero under the assumption of neutrality and is significantly negative if there is such an excess of rare alleles. Not only are there no significantly negative *D* values in the *FY* regions in the Hausa, but four of the six values are positive (table 1). Clearly, this aspect of the data does not fit simple models of directional selection, and it suggests that much of the variation does not represent new mutations arising after the fixation of *FY*O* but, instead, is evidence of old variation that has survived the sweep, via crossing-over, gene conversion, or recurrent mutation. This finding is consistent with the observation, which has been reported elsewhere (Hamblin and Di Rienzo 2000), of the *FY*O* allele on two distinct haplotypes defined by polymorphic sites on both sides of the mutation.

If, because of recombination, linked neutral variants remain polymorphic after a complete selective sweep, the frequency spectrum that results from a recent selective fixation may be quite different from that produced under a no-recombination model. If the ancestral allele can be inferred for each polymorphic site (e.g., on the basis of the outgroup sequence), a bipartite spectrum of allele frequencies may be observed—that is, an excess of both low-frequency and high-frequency nonancestral alleles (Fay and Wu 2000). The *H* statistic (Fay and Wu 2000) measures the difference between π and θ_H, a measure of diversity weighted by the homozygosity of non-

Table 3

***P* Values of HKA Tests of *FY* Regions versus Locus Pairs Regions**

	FY Region III			*FY* Region V
Locus Pair[a]	Chinese	Hausa	Italians	Chinese
1	.774	.022	.736	.294
2	.108	.023	.793	.130
3	.107	.013	.569	.130
4	.085	.059	.865	.118
5	.028	.048	.488	.080
6	.030	.003	.793	.083
7	.048	.054	.776	.098
8	.001	.001	.529	.050
10	.000	.000	.178	.072

[a] Region 9 was not included in these tests, because orangutan sequence is not available.

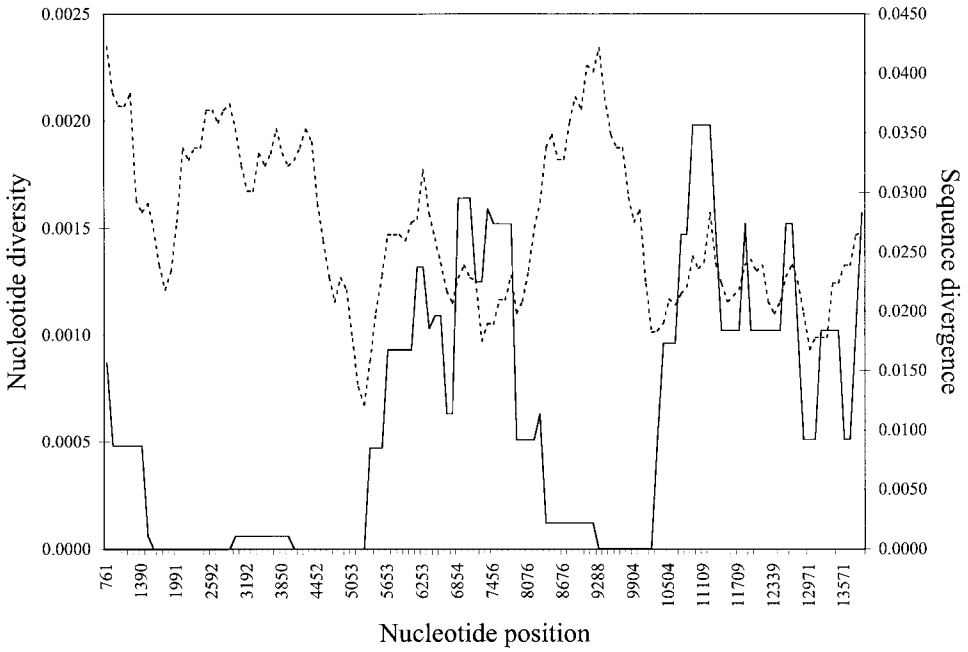

Figure 4 Sliding window of polymorphism and divergence in the Chinese sample. Regions III–VII are represented. The solid line indicates nucleotide diversity; the dashed line indicates sequence divergence from orangutan. The window size is 1,000 bp.

ancestral variants. A significant *H* statistic indicates an excess of high-frequency nonancestral alleles, consistent with the effects of hitchhiking with recombination. The *H* statistic is significant for regions III, V, and VII in the Hausa and is also significant ($P = .01$) for the entire Hausa data set (table 1). As would be expected if directional selection acted on the *FY*O* allele, this effect is observed on both sides of this site: the *H* statistic is significant ($P = .045$) also for the region III subset immediately 5′ to the *FY*O* mutation.

The Signature of Natural Selection outside Africa: the Chinese Sample

As in the case of *FY*O*, the high F_{ST} value for the *FY*A* allele may be the consequence of selective pressures acting in the populations in which this allele reaches near-fixation frequency. To explore this possibility, we collected data in a sample from a population, the Chinese, in which the *FY*A* allele is nearly fixed. Our approach to the detection of the signature of natural selection is similar to that used for the *FY*O* allele in the Hausa; however, there are two important differences. First, because the *FY*A* allele does not have any known functional consequence, it cannot be assumed that directional selection acted on this site, as opposed to one tightly linked to it. Second, because the locus-pairs data in the non-African samples showed several departures from the neutral-equilibrium model (see Discussion section), our assessment of the signature of natural selection in these samples must rely more heavily on the empirical comparisons to the locus-pairs data than on the critical values of standard neutrality tests.

In regions III and V of the Chinese sample, polymorphism is particularly low; and in region III, which spans the *FY*A* mutation, the reduction is almost significant (table 1). Because 5 of 10 pairwise tests of region III versus the locus pairs yielded *P* values <.05, variation at the *FY* gene is likely to be inconsistent with a neutral model (table 3). Pairwise tests of region V also produce consistently low *P* values, although none are significant (table 3). Because the average level of variation (across the genome) in the Chinese sample is lower than that in the Hausa (Frisse et al. 2001), the power to detect a significant reduction is correspondingly lower. Additional sequence data (regions IV and VI) were collected for the Chinese sample to increase the power of the test, but variation in these regions is not reduced (fig. 4; table 1). Thus, although regions III and V appear to have reduced variation, their interruption by a region of normal variation produces a pattern that may not be consistent with a single selective event. It has been shown that "valleys" in levels of sequence variation along a recombining chromosome may occur under the assumption of neutrality (Kim and Stephan, in press). These valleys become deeper and wider after a selective sweep, but there is a large amount of stochastic noise in the patterns, especially when N_e is small.

Although statistical evidence of a reduction of variation

in the Chinese is only marginal, the frequency spectrum and the haplotype structure of variation in this sample are clearly unusual, particularly in regions VI–VIII. These regions all have an excess of intermediate-frequency variants, which leads to significantly positive *D* statistics. In fact, if regions III–VIII are pooled, the *D* statistic is significantly high ($D = 1.78$) across this entire 23-kb region. The change in frequency spectrum (i.e., the variance in *D*) across all regions surveyed is significantly greater ($P = .022$) than would be expected under an equilibrium model that assumes free recombination between these regions. However, the variance of *D* across the locus pairs in the Chinese sample is also significantly larger than would be expected under an equilibrium model, probably as a result of different population histories. Interestingly, the variance of *D* across the eight *FY* regions is 2.42, which is substantially larger than the variance of the locus pairs (1.72). Because the 8 *FY* regions are linked, we would expect, as a result of correlated history, a *smaller* variance than that found across the 10 unlinked regions. Thus, the magnitude of the variance in *D* across a 34-kb region is unexpected, even in the context of the large variance observed at neutral loci.

As we have seen for the Hausa data, estimates of ρ are lower at *FY* than the estimates for the locus-pairs data (table 2). This implies high LD levels and an apparent reduction of the effective population size, consistent with the observation of reduced variation. To detect nonrandom associations between alleles at specific positions, we estimated two statistics, $|D'|$ and r^2, that summarize information about LD between pairs of sites (see Methods section). The vast majority of $|D'|$ values in this sample are 0.75–1.0, even at sites >20 kb apart, indicating that little or no recombination has occurred during the history of this sample (fig. 5). High r^2 values reflect a strong haplotype structure, in which pairs of sites are represented by only two gametic types. Several pairs of sites have r^2 values close to 1, suggesting that the haplotype structure of variation is unusual. To examine multisite LD in more detail, we estimated haplotypes for each population, using the program PHASE (Stephens et al. 2001; see Methods). Region I was not included in this analysis, because it is far from region II and has little informative variation. The haplotype analysis (fig. 6) shows that both the high LD and the skew toward intermediate-frequency variants in the Chinese sample result from the presence of two major haplotypes that differ at 20 positions across 20 kb and occur at almost equal frequency. A sliding-window test of haplotype number (Andolfatto et al. 1999) reveals that the number of different haplotypes present in the Chinese sample in regions II–VII is significantly less than would be expected ($P = .01$) under a standard neutral-equilibrium model with recombination.

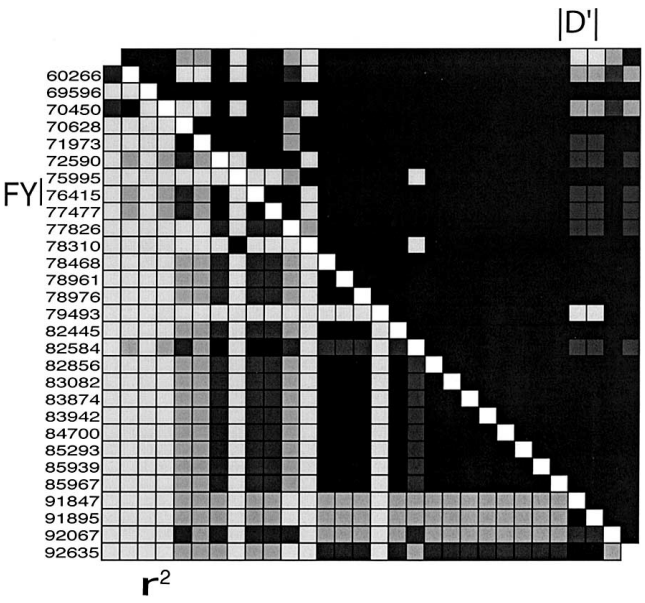

Figure 5 LD in regions II–VIII of the Chinese sample. $|D'|$ (*top*) and r^2 (*bottom*) were estimated for pairs of sites in diploid data. The shading of squares indicates the value of the statistic: blackened squares = .75–1.0; dark shaded squares = .5–.75; medium shaded squares = .25–.5; and light shaded squares = 0–.25. The numbers indicate the position of the polymorphic site. The position of the *FY* gene is marked by a bar (*left*).

The Signature of Natural Selection outside Africa: the Italian Sample

Several unusual patterns of variation are observed among the Italian sample, in the region 3′ to the *FY* gene, most notably in region V. In this region, there are 10 segregating sites and a θ_w of 0.2%, twice the mean value for humans (table 1). This high level of polymorphism results from the presence of nine polymorphic sites (eight SNPs and one single-base deletion) found in the heterozygous state clustered within a 600-bp region in a single individual (Italian subject 12; fig. 6). This individual (an *FY*A/FY*B* heterozygote) has no remarkable variation in any of the other regions surveyed. All nine polymorphisms are singletons in a region that otherwise has little variation in either the Italian or the worldwide sample. Sequencing of this region in two progeny of Italian subject 12 confirmed that all nine variants are on the same chromosome. We screened a larger collection of Italian DNA samples and found a second occurrence of this diverged haplotype, in a total of 104 chromosomes. The possibility that this haplotype is a polymorphic duplication, rather than an allele, was eliminated by probing a genomic Southern blot: no additional bands hybridized in the lanes containing DNA from the two individuals with the exceptional haplotype (data not shown).

These results are unusual for two main reasons. First,

the depth of the genealogy is very great in comparison with any other locus in humans. The amount of sequence difference between the diverged chromosome and the others in the 600-bp segment is >1%. The sequence divergence for this small segment (defined post hoc to estimate the maximum sequence divergence) is on the order of the average divergence between human and chimpanzee. (The nine mutations are evenly distributed between the two deep branches of the genealogy, as expected.) Second, there is almost no differentiation in the remaining portion of the genealogy: 109 of the 128 chromosomes surveyed in the four population samples are identical.

The presence of this divergent chromosome leads to a number of significant departures when tests of neutrality are applied (table 1). *D* is significantly negative; however, because all eight singleton mutations are on one chromosome, these low-frequency alleles have not arisen after a selective sweep. Furthermore, the HKA test detects an excess of segregating sites, rather than a deficiency. There is also an excess of high-frequency derived alleles in this region, as detected by the *H* test.

Another unusual pattern of variation 3′ to the *FY* gene occurs in region VII. As in the Chinese sample, a significantly positive *D* is observed, indicating an excess of intermediate-frequency variants. Haplotype diversity is also low, with several polymorphic sites showing a two-haplotype structure. This pattern is reminiscent of, although not as extreme as, that seen in the Chinese sample. The presence of both a significantly negative and a significantly positive *D* at two loci in the same sample is unexpected ($P = .01$), even for two unlinked regions, and is very unlikely for loci <4 kb apart (i.e., recombination on the order of 10^{-5} per generation). The variance of *D* was not unusually large in the Italian locus-pairs data.

Discussion

The Duffy blood group gene shows an extreme degree of between-population differentiation of allele frequency. We surveyed sequence variation in the *FY* region in four population samples that harbor different proportions of the three major *FY* alleles. Our goals were to characterize the signature of directional selection on *FY*O* in sub-Saharan Africa and to understand the extent to which natural selection has also played a role in the extreme geographic differentiation of the other derived allele at this locus, *FY*A*. Several unusual patterns in different aspects of sequence variation were observed, leading to a number of striking departures from the theoretical expectations of a neutral-equilibrium model as well as from empirical patterns observed at putatively neutral loci that were surveyed in the same population samples. Although the departures observed in the African sample are consistent with a history of directional selection involving recombination, those observed in the Italian and Chinese samples do not, collectively, fit a simple model of selection. These patterns may represent a more complex and previously unrecognized signature of positive selection.

The Signature of Selection on FY*O *in the Hausa*

The *FY* region in the Hausa, a sub-Saharan African population fixed for the *FY*O* allele, shows evidence of directional selection in two independent properties of the data: the level of sequence variation and the frequency spectrum. The combination of significantly low sequence variation around the *FY*O* mutation and a significant excess of high-frequency derived alleles at linked sites constitutes a pattern that would be expected after a recent episode of directional selection involving limited recombination (Fay and Wu 2000). Such a pattern was observed at the *Acp26Aa* gene, which is located in a region of high recombination in *Drosophila melanogaster*. Patterns of polymorphism and divergence at silent and replacement sites in this gene indicate that positive selection has driven the fixation of many amino acid differences (Aguadé et al. 1992; Tsaur and Wu 1997); yet there is no reduction of variation at *Acp26Aa,* as might be expected after a recent selective sweep (Tsaur et al. 1998). Instead, there are a large number of derived alleles at high frequency, a pattern that may result when selection is sufficiently weak to allow recombination before fixation. Subsequent analysis of these data, as well as a theoretical treatment of the problem (Fay and Wu 2000), showed that the greatest departure from the equilibrium-frequency spectrum is expected when *c/s* (rate of recombination per segment/selection coefficient) is ~0.01 and that variation will be completely eliminated only when *c/s* is $<2 \times 10^{-3}$. Assuming that the *FY*O* mutation is the target of selection—and given 1.4×10^{-8} per generation in the *FY* region as the rate of recombination between adjacent sites—we can estimate the selection coefficient on the *FY*O* mutation, as follows: The two polymorphisms closest, on either side, to the *FY*O* site in the Hausa are 535 bp apart and define the "footprint" of selection (i.e., the segment in which variation has been completely eliminated). The *c* value for this segment can be calculated as $(535 \text{ bp}/2) \times (1.4 \times 10^{-8}/\text{bp}) = 3.75 \times 10^{-6}$. Thus, for *c/s* values $<2 \times 10^{-3}$, *s* must be ~0.002 or larger. The significant excess of high-frequency derived alleles in regions V and VII—which are ~5 kb and 10 kb from the *FY*O* mutation, respectively—but not in region VIII, is consistent with this estimate of *s*. Note that this calculation is based on a haploid model of selection. Because the known phenotype (resistance to *P. vivax* malaria) of the *FY*O* mutation is recessive, the dynamics of the fixation pro-

CH1A
CH1B
CH2B
CH15A
CH15B
CH9B
CH3A
CH16A
CH11B
CH10A
CH4A
CH4B
CH5A
CH6A
CH6B
CH8A
CH8B
CH2A
CH14A
CH13A
CH5B
CH3B
CH7B
CH13B
CH11A
CH7A
CH9A
CH12A
CH12B
CH16B
CH10B
CH14B
H7MB
H10MB
H3FB
H4MA
H7FA
H17MA
H2MB
H3FA
H6MB
H22MA
H15MA
H2MA
H4MB
H6MA
H7FB
H7MA
H11FB
H14MA
H17MB
H19MA
H1MB
H1MA
H19MB
H11FA
H14MB
H15MB
H5FA
H10MA
H22MB
H1FA
H5FB
H1FB

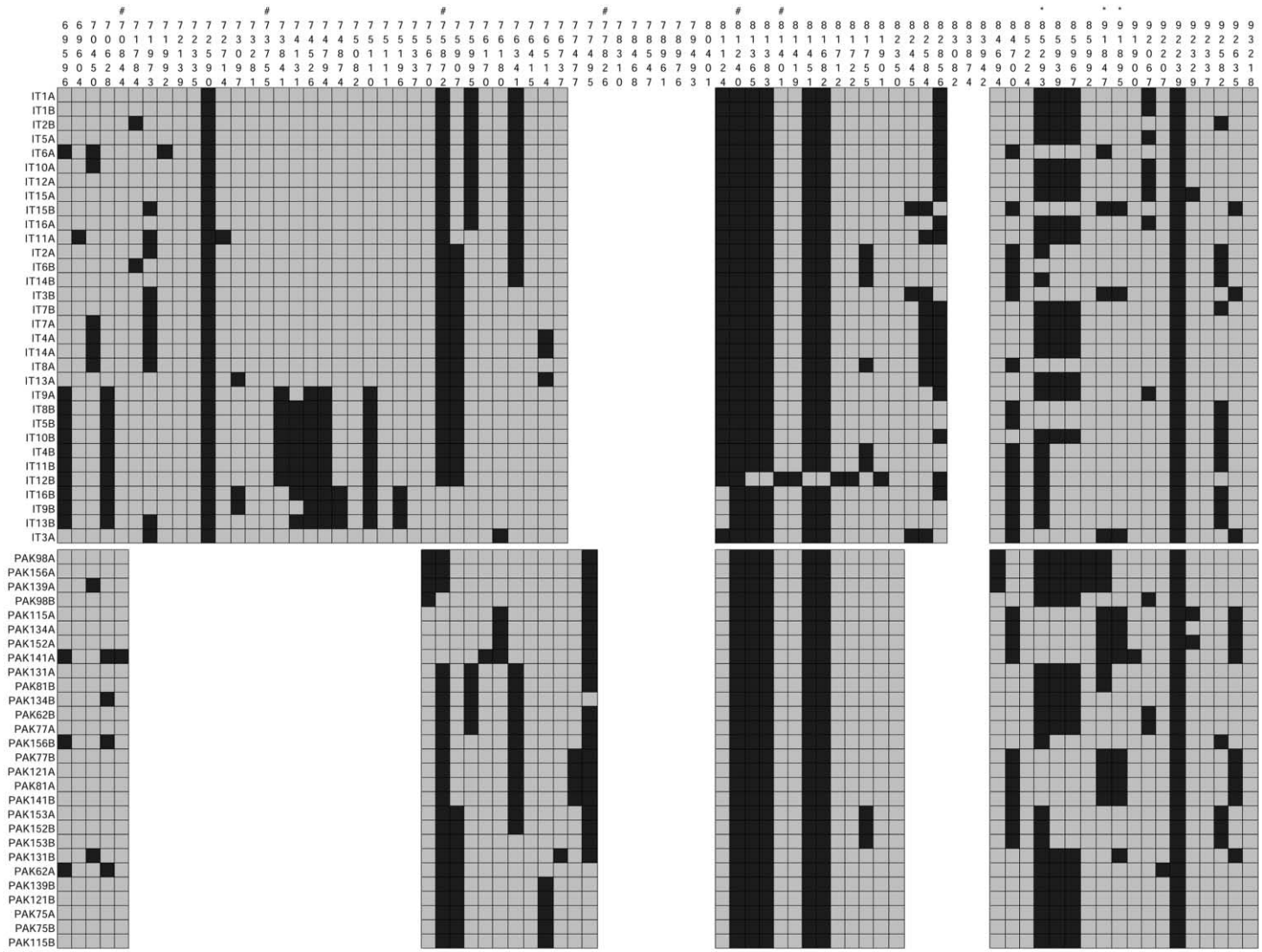

Figure 6 Estimated haplotypes for regions II to VIII. The two haplotypes inferred from one diploid genotype are labeled A and B (e.g., IT1A and IT1B). Haplotypes within a population sample are grouped to emphasize haplotype structure. In most cases, dark gray indicates a derived allele, and light gray indicates an ancestral allele. An asterisk (*) indicates that the ancestral state is unknown, in which case the color is arbitrary. A hatch (#) indicates that the ancestral state at CpG sites is ambiguous, in which case C was assumed to be ancestral. White indicates that the site was not scored in the sample.

cess would differ from those assumed in this model and would probably depend on some level of inbreeding in small population groups.

Recent theoretical work has shown that the power of the *H* test to detect selection is good when the selective fixation of the favored allele is recent (i.e., $\leq 0.4\ N_e$ generations) but declines rapidly thereafter (Przeworski 2001). It has been shown that models of population structure can also lead to statistically significant results in *H* tests. We calculated the *H* statistic for the locus-pairs data and found a significant result at 4 of 10 locus pairs in the Chinese sample. One of these four locus pairs also shows a significant departure in the Hausa and Italian samples. Significant results in *H* tests have also been obtained for a number of other data sets (Przeworski 2001), which strongly suggests that aspects of human population history contribute to these results. However, in the case of *FY*O,* four different considerations support the idea that the significant *H* tests in the Hausa sample result from directional selection, rather than from population history. First, there is a strong prior hypothesis of selection based on both allele-frequency data and functional evidence. Second, the strongest departure in the *H* statistics is observed at a distance that is consistent with the selection coefficient estimated on the basis of the footprint of selection. Third, the estimated time of fixation of *FY*O* (33,000 years, or 0.16 N_e generations) (Hamblin and Di Rienzo 2000) is within the interval of time, since the selective sweep, during which the *H* test remains significant (namely, 0.4 N_e generations). Fourth, the *H* statistic is significant in segments on either side of the putative selected (*FY*O*) site.

In the Hausa, we observed a signature of directional selection that meets fairly well the predictions of a simple model with recombination. Importantly, however, if we had not known the exact nucleotide responsible for the *FY*O* phenotype or if this were a region being surveyed at random, we would have been less likely to draw a strong conclusion of selection at this locus. Furthermore, other features of the data that may be related to a history of directional selection on *FY*O* (e.g., increased LD) might not be interpreted as such if they were observed in a random survey. In this sense, the signature of selection is not easy to detect.

The Signature of Selection in Non-African Populations

In addition to the departures observed in the Hausa sample, a number of unusual features were observed in the Chinese and Italian samples, particularly 3′ to the *FY* gene. Both of these population samples, however, appear to violate assumptions of the neutral-equilibrium model at putatively neutral loci. Estimates of N_e based on ρ and on θ were different from each other for both the Chinese and Italian locus-pairs samples (table 2). (There was no evidence of an increased variance in θ_w, i.e., no significant heterogeneity in estimates of θ_w and no departure in a multilocus HKA test.) In addition, the frequency spectrum of variation in both these samples is inconsistent with a neutral-equilibrium model: the average *D* is too large in the Italian sample, and the variance of *D* is too large in the Chinese sample (Frisse et al. 2001). Of the 10 regions, 4 have a significant ($P < .05$) *H* statistic in the Chinese sample. Thus, there are no formal neutrality tests that would be appropriate for these populations.

Nevertheless, comparisons between the locus-pairs data and the *FY* data show that the latter are even more extreme than the former, which raises the possibility that this genomic region has also been the target of selection outside Africa. Two features are clearly exceptional: the presence of a very divergent lineage in the Italian sample and the large variance of *D* in the Chinese sample. The remarkable time depth of the genealogy in the Italian sample suggests the action of evolutionary forces that can maintain genetic variation longer than would be expected under a neutral-equilibrium model. Examples of such forces are long-standing balancing selection and ancient admixture.

Region V in the Chinese sample is invariant but is flanked on both sides by regions of average variation and strongly positive *D*. Across all regions, the variance of *D* is also high, even in comparison with the large variance observed at the putatively neutral locus pairs (40% higher than at the locus pairs). Coincident with the significantly positive *D* statistics is a striking two-haplotype structure. Although the Chinese locus-pairs regions, collectively, showed a departure from equilibrium, no one region consistently appeared in the extremes of the distribution for geographic differentiation, frequency spectrum, LD, and level of variation, as the *FY* region does. These departures from neutral expectations and from the empirical patterns at the locus pairs again raise the possibility that positive natural selection underlies our observations, although they cannot be reconciled with a single simple model of selection. If selection indeed has acted on this region, it is possible that the complexity of this signature results from the nonequilibrium history of non-African populations; for example, limited theoretical (Slatkin and Wiehe 1998) and empirical (Chen et al. 2000; Ford 2000; Stephan et al. 1998) studies have shown that a variety of outcomes are possible when selection occurs in geographically structured populations. Another possibility is that multiple advantageous mutations have arisen at different sites in the regions surveyed. The advantage conferred by these variants may have changed because of environmental changes—possibly as a result of exposure to different pathogens—so that some of the patterns observed today may be the relics of older selective pressures. If multiple

advantageous variants indeed have arisen, some may have evolved by directional selection and others by balancing selection. The opposite effects on patterns of variation expected under these two simple models might also underlie the complexity of our observations.

Comparisons with Other Loci Implicated in Disease Resistance

Malaria due to *P. falciparum* has been a strong selective agent in human evolution, leading to high-frequency advantageous mutations at the β-globin, α-globin, and G6PD genes. These mutations differ from the *FY*O* mutation in that they are deleterious in the absence of malarial selection and are therefore found as balanced polymorphisms occurring at frequencies that are correlated with the local incidence of malaria. Nonetheless, because malaria-resistance alleles appear to be young, they, like *FY*O*, may show the low haplotypic diversity characteristic of an allele that has recently risen in frequency under the pressure of positive selection (e.g., see Tishkoff et al. 2001). Although the Hb S, G6PD A−, and Med alleles are typically defined by single mutations, further study of the haplotypic structure of these resistance alleles suggests that, at least in some cases, multiple mutations are involved in the phenotype. This is clearly the case for G6PD A−, in which two mutations act synergistically to produce the deficient phenotype (Town et al. 1992). Experiments in a bacteria-expression system have shown that the Val68Met mutation that distinguishes the A− allele from the A allele has only a small effect on activity when introduced into a B background. Only together with the Asn126Asp mutation, which defines the A allele, is there a large reduction of enzyme activity. The Hb S allele is found on several distinct haplotype backgrounds that are geographically restricted, and it is not clear whether this is the result of gene conversion or recurrent mutation (Flint et al. 1993). However, the genetic variation found on the different Hb S haplotypes may be important in determining the severity of sickle cell disease; for example, the Senegal Hb S haplotype is associated with an increased level of fetal hemoglobin in adult life, making it more likely that individuals bearing this haplotype will reach reproductive age (Powars et al. 1989).

An unknown pathogen is likely to have led to the increase of the CCR5-Δ32 mutation, which fortuitously confers resistance to HIV infection. Disease progression in HIV infection (Martin et al. 1998), as well as promoter strength in vitro (Mummidi et al. 2000), is affected by several polymorphisms in the *cis*-regulatory region of CCR5, which is in LD with the Δ32 mutation. If more than one mutation at a locus contributes to a selected phenotype, the dynamics of the selective response must necessarily be complex. This situation may be the rule, rather than the exception.

Target of Selection

Positive selection requires a target, but no known or predicted genes are located in any of the sequenced regions 3′ to the *FY* gene. We also failed to find any cDNAs that matched the sequence in this region. However, important regulatory elements can be located 3′ to a gene. Because interspecific comparisons have proved useful for identifying conserved noncoding sequences that may have a regulatory role (e.g., see Loots et al. 2000), we compared the orthologous mouse and human sequences. Several conserved noncoding sequences (>70% identical) were identified that were consistent with the presence of a regulatory element such as an enhancer. A similar pattern of sequence conservation was found, for example, 3′ to the *H19* gene and was shown to be associated with the presence of novel tissue-specific enhancers (Ishihara et al. 2000). None of the conserved elements 3′ to the *FY* gene were located in the region of low variation of region V. Thus, it is unclear whether this region has functional significance. Two of the intermediate-frequency variants found in the Chinese sample (positions 78961 and 83082) are located in conserved noncoding sequences, as might be expected if one or more of them were the target of recent balancing selection. In addition, not all conserved sequences were resequenced in the population samples, so one of them might contain the true target of selection. In vitro or transgenic assays will be necessary to determine whether a functional element and potential target of selection exists 3′ to the *FY* gene.

Speculation on the nature of selection on *FY*-gene expression is limited by our lack of understanding of the function of its gene product, Duffy antigen receptor for chemokines (DARC). DARC differs from other chemokine receptors in that it binds chemokines of both the C-C and C-X-C types and does not appear to induce signal transduction (Hadley and Peiper 1997). Its function in red blood cells, where it serves as a receptor for *P. vivax,* is clearly not essential, since the vast majority of sub-Saharan Africans lack it. However, DARC is also found on endothelial cells in numerous tissues throughout the body and in Purkinje cells of the cerebellum. It has been proposed that DARC serves as a sink for chemokines and may have a role in regulating the inflammatory response. Up-regulation of DARC levels has been observed in renal cells of patients with renal disease (Liu et al. 1999), and differences in response to lipopolysaccharide injection were observed in mice homozygous for a deletion of the *Dfy* gene (Dawson et al. 2000; Luo et al. 2000). Susceptibility to bacterial infection was not affected in the knockout mice, however,

and they appeared to be healthy and developmentally normal, as are the rare human beings who completely lack DARC (Hadley and Peiper 1997).

The apparent health of mice and humans that lack DARC does not preclude an important role for this protein, under certain environmental conditions. Total elimination of DARC expression in response to selective pressure by *P. vivax* could have been accomplished by a large number of coding mutations, yet the *FY*O* mutation occurs within a 6-bp GATA element and results in an erythroid-specific phenotype. This suggests that there may have been selective pressure to preserve DARC function in nonerythroid cells; directional selection could act on some aspect of this function. Alternatively, considering the numerous examples of chemokine receptors that are exploited by pathogens (Pease and Murphy 1998), it is possible that other pathogens have exerted selective pressure on DARC.

Conclusions

Both derived alleles of the Duffy blood group locus—*FY*O* and *FY*A*—have striking patterns of geographic differentiation, resulting in values of F_{ST} that are among the highest observed in human populations and strongly suggesting the action of natural selection. The *FY*O* allele in sub-Saharan Africans represents a best-case scenario for detecting the effects of directional selection on patterns of sequence variation: the phenotype is clear and well understood, the precise nucleotide location of the responsible mutation is known, and the population is more variable and closer to equilibrium than are non-African populations. Indeed, standard tests of neutrality have enabled us to detect a clear signature of natural selection surrounding the *FY*O* mutation. In contrast, our study of the *FY*A* allele in non-African populations required us to take an empirical approach, because of the lack of equilibrium in our study populations. We find evidence that evolution in the *FY* region has been unusual in both Italian and Chinese samples but that simple models cannot explain the data. The complex patterns of variation that we observe may result either from the interaction of selection with population history or from the interaction of multiple selective forces acting on more than one target.

Acknowledgments

We are especially grateful to C. Langley for his enthusiasm for this project and his thoughtful comments on previous versions of the manuscript and to J. Hey for many helpful discussions during the early stages of this project. We thank J. Donfack, G. Galluzzi, W.-H. Li, and A. Novelletto, for providing samples; Matthew Stephens, for providing the PHASE program; and F. Depaulis, J. Fay, M. Fullerton, R. Hudson, and M. Przeworski, for discussions and advice. We are grateful to two anonymous reviewers for their helpful comments. This work was supported by National Human Genome Research Institute award HG02098 to A.D.R. E.E.T. was supported by a predoctoral training grant from the National Institutes of Health.

Electronic-Database Information

Accession numbers and URLs for data in this article are as follows:

GenBank, http://www.ncbi.nlm.nih.gov/Genbank/ (for accession number AL035403)
Hudson Laboratory, http://home.uchicago.edu/~rhudson1/ (for a program to estimate ρ)
Online Mendelian Inheritance in Man (OMIM), http://www.ncbi.nlm.nih.gov/Omim/ (for *FY* [MIM 110700])

References

Aguadé M, Miyashita N, Langley CH (1992) Polymorphism and divergence in the Mst26A male accessory gland gene region in *Drosophila*. Genetics 132:755–770

Andolfatto P, Wall JD, Kreitman M (1999) Unusual haplotype structure at the proximal breakpoint of In(2L)t in a natural population of Drosophila melanogaster. Genetics 153:1297–1311

Berry A, Kreitman M (1993) Molecular analysis of an allozyme cline: alcohol dehydrogenase in *Drosophila melanogaster* on the east coast of North America. Genetics 134:869–893

Bowcock AM, Kidd JR, Mountain JL, Hebert JM, Carotenuto L, Kidd KK, Cavalli-Sforza LL (1991) Drift, admixture, and selection in human evolution: a study with DNA polymorphisms. Proc Natl Acad Sci USA 88:839–843

Braverman JM, Hudson RR, Kaplan NL, Langley CH, Stephan W (1995) The hitchhiking effect on the site frequency spectrum of DNA polymorphisms. Genetics 140:783–796

Cavalli-Sforza LL (1966) Population structure and human evolution. Proc R Soc Lond B Biol Sci 164:362–379

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

Chen Y, Marsh BJ, Stephan W (2000) Joint effects of natural selection and recombination on gene flow between *Drosophila ananassae* populations. Genetics 155:1185–1194

Dawson TC, Lentsch AB, Wang Z, Cowhig JE, Rot A, Maeda N, Peiper SC (2000) Exaggerated response to endotoxin in mice lacking the Duffy antigen receptor for chemokines (DARC). Blood 96:1681–1684

Fay JC, Wu CI (2000) Hitchhiking under positive Darwinian selection. Genetics 155:1405–1413

Flint J, Harding RM, Clegg JB, Boyce AJ (1993) Why are some genetic diseases common? distinguishing selection from other processes by molecular analysis of globin gene variants. Hum Genet 91:91–117

Ford MJ (2000) Effects of natural selection on patterns of DNA sequence variation at the transferrin, somatolactin, and p53 genes within and among chinook salmon (*Oncorhynchus tshawytscha*) populations. Mol Ecol 9:843–855

Frisse L, Hudson RR, Bartoszewicz A, Wall JD, Donfack J, Di

Rienzo A (2001) Gene conversion and different population histories may explain the contrast between polymorphism and linkage disequilibrium levels. Am J Hum Genet 69:831–843

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

Hamblin MT, Di Rienzo A (2000) Detecting the signature of natural selection in humans: evidence from the Duffy blood group locus. Am J Hum Genet 66:1669–1679

Hill WG (1974) Estimation of linkage disequilibrium in randomly mating populations. Heredity 33:229–239

Hudson RR. Two-locus sampling distributions and their application. Genetics (in press)

Hudson RR, Kreitman M, Aguadé M (1987) A test of neutral molecular evolution based on nucleotide data. Genetics 116: 153–159

Ishihara K, Hatano N, Furuumi H, Kato R, Iwaki T, Miura K, Jinno Y, Sasaki H (2000) Comparative genomic sequencing identifies novel tissue-specific enhancers and sequence elements for methylation-sensitive factors implicated in Igf2/H19 imprinting. Genome Res 10:664–671

Karl SA, Avise JC (1992) Balancing selection at allozyme loci in oysters: implications from nuclear RFLPs. Science 256: 100–102

Kim Y, Stephan W. Detecting a local signature of genetic hitchhiking along a recombining chromosome. Genetics (in press)

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

Lewontin RC, Krakauer J (1973) Distribution of gene frequency as a test of the theory of the selective neutrality of polymorphisms. Genetics 74:175–195

Liu XH, Hadley TJ, Xu L, Peiper SC, Ray PE (1999) Upregulation of Duffy antigen receptor expression in children with renal disease. Kidney Int 55:1491–1500

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

Loots GG, Locksley RM, Blankespoor CM, Wang ZE, Miller W, Rubin EM, Frazer KA (2000) Identification of a coordinate regulator of interleukins 4, 13, and 5 by cross-species sequence comparisons. Science 288:136–140

Luo H, Chaudhuri A, Zbrzezna V, He Y, Pogo AO (2000) Deletion of the murine Duffy gene (*Dfy*) reveals that the Duffy receptor is functionally redundant. Mol Cell Biol 20: 3097–3101

Martin MP, Dean M, Smith MW, Winkler C, Gerrard B, Michael NL, Lee B, Doms RW, Margolick J, Buchbinder S, Goedert JJ, O'Brien TR, Hilgartner MW, Vlahov D, O'Brien SJ, Carrington M (1998) Genetic acceleration of AIDS progression by a promoter variant of CCR5. Science 282:1907–1911

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

Mummidi S, Bamshad M, Ahuja SS, Gonzalez E, Feuillet PM, Begum K, Galvis MC, Kostecki V, Valente AJ, Murthy KK, Haro L, Dolan MJ, Allan JS, Ahuja SK (2000) Evolution of human and non-human primate CC chemokine receptor 5 gene and mRNA. Potential roles for haplotype and mRNA diversity, differential haplotype-specific transcriptional activity, and altered transcription factor binding to polymorphic nucleotides in the pathogenesis of HIV-1 and simian immunodeficiency virus. J Biol Chem 275:18946–18961

Nachman MW (2001) Single-nucleotide polymorphisms and recombination rate in humans. Trends Genet 17:481–485

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

Powars DR, Chan L, Schroeder WA (1989) The influence of fetal hemoglobin on the clinical expression of sickle cell anemia. Ann N Y Acad Sci 565:262–278

Przeworski M. The signature of natural selection at randomly chosen loci. Genetics (in press)

Rozas J, Rozas R (1999) DnaSP version 3: an integrated program for molecular population genetics and molecular evolution analysis. Bioinformatics 15:174–175

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

Stephan W, Xing L, Kirby DA, Braverman JM (1998) A test of the background selection hypothesis based on nucleotide data from *Drosophila ananassae*. Proc Natl Acad Sci USA 95:5649–5654

Stephens M, Smith NJ, 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

Taylor MF, Shen Y, Kreitman ME (1995) A population genetic test of selection at the molecular level. Science 270:1497–1499

Tishkoff SA, Varkonyi R, Cahinhinan N, Abbes S, Argyropoulos G, Destro-Bisol G, Drousiotou A, Dangerfield B, Lefranc G, Loiselet J, Piro A, Stoneking M, Tagarelli A, Tagarelli G, Touma EH, Williams SM, Clark AG (2001) Haplotype diversity and linkage disequilibrium at human G6PD: recent origin of alleles that confer malarial resistance. Science 293:455–462

Tournamille C, Colin Y, Cartron JP, Le V, 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

Town M, Bautista JM, Mason PJ, Luzzatto L (1992) Both mutations in G6PD A− are necessary to produce the G6PD deficient phenotype. Hum Mol Genet 1:171–174

Tsaur SC, Ting CT, Wu CI (1998) Positive selection driving the evolution of a gene of male reproduction, Acp26Aa, of *Drosophila*. II. Divergence versus polymorphism. Mol Biol Evol 15:1040–1046

Tsaur SC, Wu CI (1997) Positive selection and the molecular evolution of a gene of male reproduction, Acp26Aa of *Drosophila*. Mol Biol Evol 14:544–549