Disentangling the Effects of Demography and Selection in Human History

Jason E. Stajich and Matthew W. Hahn†*

*Department of Molecular Genetics and Microbiology, Duke University, Durham, North Carolina; and †Center for Population Biology, University of California, Davis

Demographic events affect all genes in a genome, whereas natural selection has only local effects. Using publicly available data from 151 loci sequenced in both European-American and African-American populations, we attempt to distinguish the effects of demography and selection. To analyze large sets of population genetic data such as this one, we introduce "Perlymorphism," a Unix-based suite of analysis tools. Our analyses show that the demographic histories of human populations can account for a large proportion of effects on the level and frequency of variation across the genome. The African-American population shows both a higher level of nucleotide diversity and more negative values of Tajima's *D* statistic than does a European-American population. Using coalescent simulations, we show that the significantly negative values of the *D* statistic in African-Americans and the positive values in European-Americans are well explained by relatively simple models of population admixture and bottleneck, respectively. Working within these nonequilibrium frameworks, we are still able to show deviations from neutral expectations at a number of loci, including *ABO* and *TRPV6*. In addition, we show that the frequency spectrum of mutations—corrected for levels of polymorphism—is correlated with recombination rate only in European-Americans. These results are consistent with repeated selective sweeps in non-African populations, in agreement with recent reports using microsatellite data.

Introduction

Uncovering the evolutionary processes that determine the frequency of alleles in a population and differences between populations or species is a major goal of population genetics. The effects of selective, demographic, and random processes at any single locus can cause deviations from the neutral-equilibrium expectation, but determining which of these forces has had the most influence on DNA variation is often difficult (e.g., Tajima 1989; Gillespie 1991; Braverman et al. 1995; Simonsen, Churchill, and Aquadro 1995; Fu 1997; Galtier, Depaulis, and Barton 2000; Hahn, Rausher, and Cunningham 2002). The use of large, multilocus data sets from the same individuals in natural populations offers the opportunity to study the combined effects of these evolutionary processes across the genome. By using a large number of loci, we may more easily be able to disentangle the genome-wide effects of demography from the locus-specific effects of natural selection (e.g., Begun and Whitley 2000; Wall, Andolfatto, and Przeworski 2002; Glinka et al. 2003).

By far, the largest data sets on sequence variation have been collected for humans (Cargill et al. 1999; Halushka et al. 1999; Sachidanandam et al. 2001). Using this huge amount of data, researchers have been able to develop a clearer picture of human demographic history and the rise of anatomically modern humans from Africa (Cavalli-Sforza, Menozzi, and Piazza 1994; Jorde et al. 1997; Excoffier and Schneider 1999; Harpending and Rogers 2000; Goldstein and Chikhi 2002; Marth et al. 2003). In addition, evidence for the action of adaptive natural selection has been found despite the concurrent effects of migration, bottlenecks, and population expansion (Bowcock et al. 1991; Tishkoff et al. 1996; Akey et al. 2002; Shen et al. 2002; Rockman et al. 2003; Hahn et al. 2004).

Key words: hitchhiking, ABO, TRPV6, bottleneck, admixture.

E-mail: mwhahn@ucdavis.edu.

Mol. Biol. Evol. 22(1):63–73. 2005
doi:10.1093/molbev/msh252
Advance Access publication September 8, 2004

Large numbers of loci sampled across regions of differing recombination rates have also afforded a view of the effects of linked selection on nucleotide variation. It has been shown across a number of species that the recombination rate is correlated with levels of polymorphism but not with levels of divergence (Stephan and Langley 1989; Begun and Aquadro 1992; Nachman 1997; Dvorak, Luo, and Yang 1998; Kraft et al. 1998; Stephan and Langley 1998; Przeworski, Hudson, and Di rienzo 2000; Nachman 2001; Tenaillon et al. 2001; Payseur and Nachman 2002). Strong selection, whether advantageous ("hitchhiking" [Maynard Smith and Haigh 1974]) or deleterious ("background selection" [Charlesworth, Morgan, and Charlesworth 1993]), could be responsible for reduced genetic variability at linked loci and would create a correlation between polymorphism and recombination; this linked selection would have no effect on rates of divergence (Birky and Walsh 1988). Because hitchhiking also predicts the excess of low-frequency mutations found in regions of low recombination over a much broader range of parameters and sample sizes than does background selection (Kaplan, Hudson, and Langley 1989; Hudson and Kaplan 1994; Charlesworth, Charlesworth, and Morgan 1995; Braverman et al. 1995; Gillespie 2000), this explanation has been favored. However, it has recently been found that there is a correlation between recombination rate and both polymorphism and divergence in data from humans (Hellmann et al. 2003). This result suggests a neutral explanation for the correlation between recombination and polymorphism in humans: namely, that recombination itself is mutagenic (Hellmann et al. 2003).

To both distinguish between demographic and selective effects and to study the effects of linked selection in the human genome, we analyzed 151 loci all sequenced in the same individuals from both an African-American and a European-American population. By first examining the overall influence of demographic histories on sequence variation in each population, we were able to construct a nonequilibrium expectation against which we could test for natural selection. Because the loci are scattered across regions whose recombination rates differ by an order of

Table 1
Average Summary Statistics of Nucleotide Variation Across Loci

Population	θ[a]	π	Tajima's *D* Statistic	*P*-value[b]
African-American	0.0011	0.0009	−0.49	0.0032
European-American	0.0006	0.0007	+0.26	<0.0001

[a] Watterson's estimator of θ based on the number of segregating sites (Watterson 1975).

[b] Combined probability of *D* statistic values at all 151 loci under a neutral-equilibrium model.

magnitude, we were also able to study the influence of recombination on levels and frequency of nucleotide polymorphism in both populations. Finally, the differing demographic histories of African-Americans and European-Americans allow us to contrast the effects of natural selection between human populations.

Materials and Methods

Human Sequences

Unphased genotype data for 151 genes sequenced in 24 African-Americans (12 male/12 female) and 23 European-Americans (12 male/11 female) was obtained from the SeattleSNPs Web site (http://pga.mbt.washington.edu) on January 27, 2004 (Carlson et al. 2003, 2004; Crawford et al. 2004). These 151 genes are found on 21 chromosomes (including five genes on the X chromosome) and have 19.3 kilobases (kb) sequenced per locus on average. The regions sequenced contain mostly noncoding DNA in and around the coding region. Sex-averaged recombination rates for the regions containing each of the 151 loci are from the deCODE genetic map (Kong et al. 2002) using the August 2001 freeze of the Human Genome Project Working Draft available through the UCSC Human Genome Browser (http://genome.ucsc.edu).

Population Genetic Analysis

To analyze large sets of population genetic data, we created the "Perlymorphism" suite of analysis tools, available through the open-source Bioperl project (http://www.bioperl.org [Stajich et al. 2002]). Perlymorphism calculates standard polymorphism summary statistics including π (Tajima 1983), θ (Watterson 1975), Tajima's *D* statistic (Tajima 1989), Fu and Li's *D*, *D**, *F*, and *F** statistics (Fu and Li 1993), F_{ST} (Wright 1951), and various measures of linkage disequilibrium (e.g., Weir 1996). In addition, Perlymorphism can be used to test selective hypotheses by generating distributions of coalescent genealogies (Hudson 1990). In particular, it can calculate the probability of seeing a value equal to or more extreme than one of the summary statistics listed above, or *P* values associated with more general tests using the coalescent such as Hudson's haplotype test (Hudson et al. 1994) or Hahn, Rausher, and Cunningham's (2002) heterogeneity test. We used Perlymorphism to calculate π, θ, and Tajima's *D* statistic values. We also used Perlymorphism to calculate *P* values for the *D* statistic under a neutral-equilibrium model conditioned on the number of segregating sites with no recombination for all 151 loci. Values of π and θ for all five X-linked loci were multiplied by 4/3 to correct for a lower effective population size, assuming no sexual selection and equal mutation rates on the X chromosome. All further statistical analyses were carried out in JMP (SAS Institute, Inc.).

Population Modeling

We used the program "ms" and associated tools (Hudson 2002; Thornton 2003) to generate coalescent genealogies under a Wright-Fisher equilibrium model and a demographic model thought to closely mimic the histories of European-American and African-American populations (after Marth et al. [2003]). The model considered an ancestral population of size N_{anc} that split into two populations, designated P_0 (of size N_0) and P_1 (of size N_1), $N_0{*}T_{anc}$ generations ago. Population P_0 continued at equilibrium ($N_0 = N_{anc}$), whereas population P_1 went through a contraction immediately after the split. The contraction was considered a simple stepwise bottleneck (as in Fay and Wu [1999] and Marth et al. [2003]) of severity N_1/N_{anc} and length T_1. The difference between T_{anc} and T_1 is the time before the present that the contraction ended and population P_1 rebounded to size N_{1*}. We generated 10,000 genealogies conditioned on the average θ found in the European-American sample ($\theta = 13$) and that were carried out with population recombination parameter, $4N_er$, equal to 7 ($r = 1 \times 10^{-8}$, $N_e = 10{,}000$, 19,000 bases per locus).

Results and Discussion

The data set that we analyze here contains approximately 2.9 megabases (MB) sequenced in each of 24 African-Americans and 23 European-Americans (Carlson et al. 2003, 2004; Crawford et al. 2004). There were 13,943 segregating sites in the African-American sample and 8,339 segregating sites in the European-American sample; the segregating variants include both single-nucleotide polymorphisms (SNPs) and biallelic insertion/deletion polymorphisms. These resources, and the tools we have created to analyze them, offer an opportunity to study the population genomics of human history.

Demographic Effects

Data from 151 loci scattered throughout the human genome show that the African-American (AA) and European-American (EA) populations differ in multiple ways expected under an out-of-Africa scenario (Cann, Stoneking, and Wilson 1987). Average measures of nucleotide diversity in the African-American sample are significantly higher than those in the European-American sample (table 1) ($\theta_{AA} = 0.0011$ [±0.0003], $\theta_{EA} = 0.0006$ [±0.0003], paired *t*-test $P < 0.0001$; $\pi_{AA} = 0.0009$ [±0.0004], $\pi_{EA} = 0.0007$ [±0.0004], paired *t*-test $P < 0.0001$). We found no significant differences in the five X-linked loci and so include them in all further analyses. In the largest study to date of variation in a European population, Marth et al. (2003) found a similar average value for π across the genome: 0.00076. Frisse et al. (2001) found similar reductions to those shown here in nucleotide polymorphism in European relative to African populations. Surprisingly, given that we are comparing a European population to an admixed population of African and European alleles, our results actually show a slightly larger difference in levels of variation between populations. This difference may be caused by the much larger number of loci analyzed here (151 versus 10 in Frisse et al. [2001]) or the fact that the African alleles in the African-American sample come from a number of populations, while Frisse et al. (2001) sampled only the Hausa population.

In addition to differences in levels of nucleotide polymorphism, there are also large differences in the average value of Tajima's *D* statistic (Tajima 1989), a measure of the mutation frequency spectrum, between populations and from the neutral-equilibrium expectation in both populations (table 1 and fig. 1). Tajima's *D* statistic is positive when there is an excess of high-frequency mutations, as after a population contraction or under balancing selection (Tajima 1989). Tajima's *D* statistic is negative when there is an excess of low-frequency mutations, as after a population expansion, a recent selective sweep, weak negative selection, or, as we shall show, when a sample comes from an admixed population. The average *D* statistic value for the African-American population is −0.49 (±0.60), and for the European-American population it is +0.26 (±0.92) (paired *t*-test, $P < 0.0001$). These values appear to indicate overall deviations from the neutral-equilibrium hypothesis for both populations, although in opposite directions. To test the significance of these deviations, we first calculated one-tailed *P* values for each locus in the European-American sample testing for an excess of high-frequency mutations—as is expected after a population contraction. A combined probability test (Sokal and Rohlf 1995) shows a significant deviation from the neutral-equilibrium expectation ($\mathrm{df} = 302$, $P < 0.0001$). This indicates that the genome as a whole, as represented by 151 loci, shows a pattern consistent with a recent population contraction in Europeans. We also calculated one-tailed *P* values testing for an excess of *low*-frequency mutations in the African-American sample; a combined probability test is again significant ($\mathrm{df} = 302$, $P = 0.0032$). Although most African populations are thought to have been relatively stable (Kimmel et al. 1998; Pluzhnikov, Di Rienzo, and Hudson 2002), African-Americans represent an admixture between African and European populations (Chakraborty et al. 1992; Parra et al. 1998). We believe that admixture is the reason for the excess of negative values in the African-American population (Wakeley 2000; Ptak and Przeworski 2002), as we demonstrate with simulations in the next section.

The European-American population has both a distribution of *D* statistic values that is shifted more positively and a significantly larger variance in *D* statistic values than the African-American population ($F_{\max}$-test, $P < 0.01$ [fig. 1]). No values of the *D* statistic are greater than +2 or less than −2 in the African-American population, and only two are significant in a two-tailed test at the 0.05 level under the neutral-equilibrium hypothesis. After a Dunn-Sidak correction for 151 separate comparisons (Sokal and Rohlf 1995), none are significant at the nominal 0.05 level.

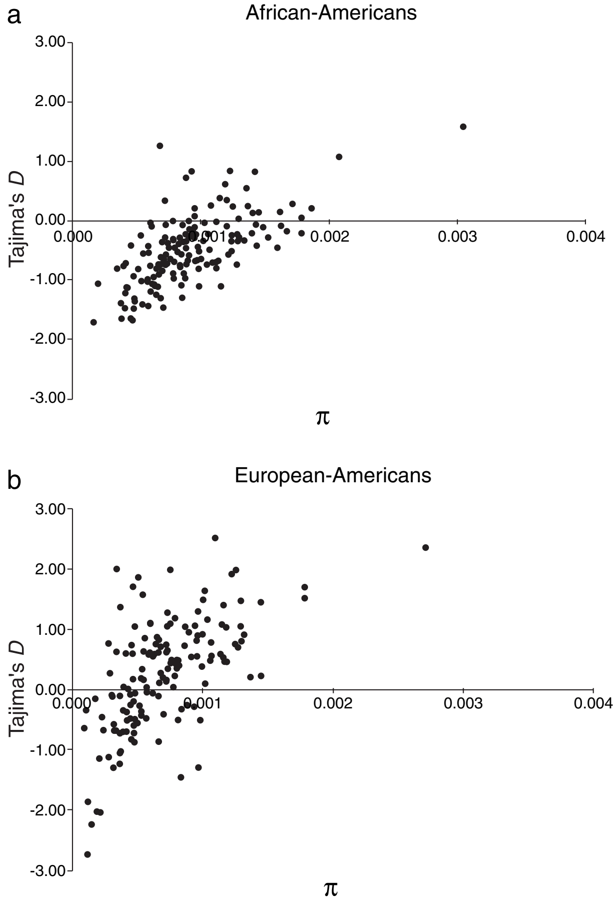

FIG. 1.—Tajima's *D* statistic and π for all 151 loci. (*a*) African-American population; (*b*) European-American population.

The European-American population, on the other hand, contains three loci with *D* statistic values greater than +2 and four loci with values less than −2 (fig. 1*b*); three of these loci remain significant at the 0.05 level after correcting for multiple comparisons (*KEL, TRPV5*, and *TRPV6* [see below]). Although an increase in average *D* statistic values is expected under a population contraction, the extreme, negative values of the *D* statistic found in this population may be caused either by repeated episodes of positive selection during adaptation to new environments or, again, by a large sampling variance during a bottleneck (Przeworksi, Wall and Andolfatto 2001). In the next section, we consider various demographic scenarios and the effects of each on values of the *D* statistic and levels of polymorphism.

Demographic Models

It is widely believed that European populations went through a severe population bottleneck during the migration of humans out of Africa and into Europe (e.g., Cann, Stoneking, and Wilson 1987; Harpending et al. 1998; Goldstein and Chikhi 2002; Marth et al. 2003, 2004). African-American populations, such as the sample we analyze here, are admixtures of multiple African and European populations (Chakraborty et al. 1992; Parra et al. 1998). To study the effects of these demographic events on

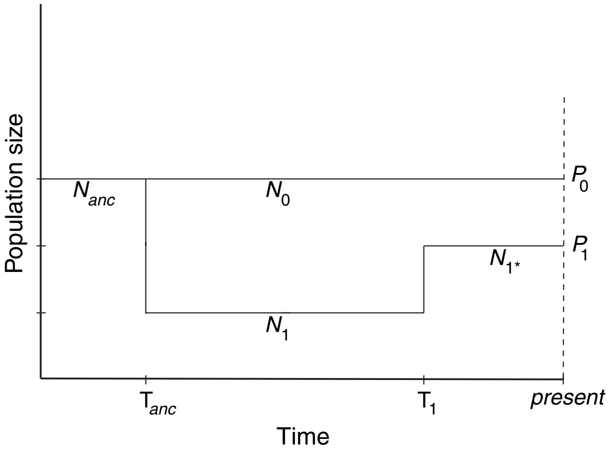

FIG. 2.—Model used for simulating human history. Populations P_0 and P_1 are formed when an ancestral population of size N_{anc} splits T_{anc} generations ago. Population P_1 has an immediate reduction in population size to N_1 that lasts until T_1 generations ago, when it rebounds to size N_{1*}.

genetic variation, we generated coalescent genealogies under conditions thought to mimic both a population bottleneck and admixture. Our aim is not to test the validity or accuracy of any particular demographic scenario, but rather to use a realistic model so that inferences about natural selection can be made. Indeed, as we show in the next section, the data used here are likely to be inappropriate for model fitting, as many of the loci may be affected by natural selection (Wall and Przeworski 2000). All of this does not mean that there is not some other model of human history that may be more or less appropriate than the one used here; in fact, any realistic model could be incorporated.

There is a surprisingly good match between relatively simple demographic models of a population bottleneck and admixture and the data we have collected. Our model (fig. 2 [Marth et al. 2003]) considered an ancestral population of size N_{anc} that split into two populations, P_0 (of size N_0; this is analogous to an African population) and P_1 (of size N_1; this is analogous to a European population), N_0*T_{anc} generations ago. The best estimate of the human effective population size (~N_0) is 10,000 (Harpending et al. 1998). If we assume a generation time of 25 years (Marth et al. 2003), then a value for $T_{anc} = 0.15$ is equivalent to a wave of migration to Europe approximately 37,500 years ago, consistent with current estimates (Harpending et al. 1998; Excoffier and Schneider 1999; Reich et al. 2001; Goldstein and Chikhi 2002; Marth et al. 2003). Coincident with this split, we envisaged a stepwise (i.e., instantaneous) population contraction only in population P_1 from size N_{anc} to size N_1. We made the severity of this reduction in size, N_1/N_{anc}, equal to one third. The length of the contraction (in N_0 generations) was $T_1 = 0.10$. Therefore, population P_1 rebounded instantaneously from size N_1 to size N_{1*}, N_0*(T_{anc}-T_1) generations ago (approximately 12,500 years ago, around the time of the discovery of agriculture [Cavalli-Sforza, Menozzi, and Piazza 1994]). The new size of this population, N_{1*}, was restored to 2/3*N_{anc} after the contraction. Genealogies conforming to this population model were generated using the program "ms" (Hudson 2002) (see *Materials and Methods* for more details). A total of 46 chromosomes, all from population P_1, were sampled for 10,000 genealogies to simulate a bottlenecked European-American population. Analogous genealogies with 48 chromosomes were generated to simulate the African-American population, except that 37 chromosomes were sampled from population P_0 and 11 chromosomes were sampled from population P_1. This admixed sample approximates the average number of African and European alleles in African-Americans individuals (Chakraborty et al. 1992; Parra et al. 1998).

We calculated Tajima's D statistic for samples from both the bottlenecked (P_1) and admixed (P_0 plus P_1) scenarios. The average D statistic value for 10,000 samples from population P_1 was +0.28 (compared with average for European-American loci: +0.26) and the average number of segregating sites (S) was 60 (compared with $S = 55$ for European-American data). The average D statistic for 10,000 admixed samples was −0.39 (compared with average for African-American loci: −0.49) and the average number of segregating sites was 100 (compared with $S = 92$ for African-American data). These values match well. In comparison, a model of an equilibrium population gives a mean D statistic value of −0.05 ($S = 57$); an admixture between two equilibrium populations gives a mean of −0.60 (S= 76); a population contraction at the same time in the past with no subsequent expansion gives a mean of +0.64 ($S = 49$); and an expansion at the same time in the past without the preceding contraction gives a mean of −0.29 ($S = 63$).

Our simple model corresponds to a history of European-American and African-American populations dominated by an out-of-Africa migration approximately 40,000 years ago (Harpending et al. 1998; Excoffier and Schneider 1999; Reich et al. 2001; Goldstein and Chikhi 2002; Marth et al. 2003) and an expansion of European populations approximately 13,000 years ago (Cavalli-Sforza, Menozzi, and Piazza. 1994; Goldstein and Chikhi 2002). Without either of these two major epochs, there remains a large proportion of unexplained data. Although there are only approximately 15,000 unique polymorphisms in the data set, our results are in accord with inferences from 500,000 SNPs analyzed in a previous study (Marth et al. 2003). Although our simple demographic model matches the observed mean Tajima's D statistic and number of segregating sites well, we should point out that this does not ensure that the variances are the same. Highly heterogeneous recombination rates among loci and the inclusion of five X-linked genes means that any simple model may not fully capture the tails of the distributions. Because of this uncertainty, we use additional, extreme examples of population histories to bolster our inferences of natural selection below.

Our modeling results lead us to two main conclusions: (1) An increased variance in D statistic values is expected after a population bottleneck, as is seen in the European-American data (Przeworksi, Wall and Andolfatto 2001). Negative values as extreme or more extreme than all but one of those observed in the data are also found from our model. These observations lead us to conclude that no special pleading for the action of natural selection is necessary to explain the majority of the

European-American data. (2) An admixed population that comes from a combination of equilibrium and bottlenecked subpopulations will have an excess of low-frequency mutations across the genome (Wakeley 2000). We were able to recapitulate the patterns seen in African-Americans by simulating such an admixture and, thus, conclude that there is no perturbing evolutionary force other than mating that is responsible for the overall deviation in the frequency spectrum we see in the data. Although other demographic models may be slightly more realistic and may fit the data slightly better, we show in the next section that the frequency of mutations at many loci has been independently affected by natural selection and, thus, are not appropriate for fitting by a neutral demographic model.

Selective Effects

Although a nonequilibrium demographic model of human history matches the observed patterns of genetic variation very well, we also found both general signatures and specific instances of adaptive natural selection throughout the genome.

Effects of Hitchhiking Across the Genome

Rather than readdressing the issue of whether polymorphism and divergence are both correlated with recombination rates in humans, we attempted to use a separate prediction of the hitchhiking model to ask whether this type of selection plays a role in shaping the level and frequency of polymorphism across the genome. The relationship between the mutation frequency spectrum and recombination expected under the hitchhiking model will not be expected under any neutral model, except because of sampling error (Braverman et al. 1995; Andolfatto and Przeworski 2001; Payseur and Nachman 2002). Because summary statistics of the frequency spectrum (such as Tajima's D statistic) use measures of nucleotide diversity (such as π and θ) in their calculations, a positive correlation between the frequency spectrum and polymorphism can be induced by sampling error (Tajima 1989). Therefore, to test for a relationship between the frequency spectrum and recombination, we must first control for any correlations between polymorphism and recombination. Because the estimated recombination rates for the regions in this study are not normally distributed, we log-transformed all values. We found a significant, but weak, relationship between both π and θ and recombination for both the African-American and European-American data sets (fig. 3) (π_{AA}, $R^2 = 0.07$, $F = 11.6$, $P = 0.0008$; θ_{AA}, $R^2 = 0.04$, $F = 6.6$, $P = 0.011$; π_{EA}, $R^2 = 0.12$, $F = 20.5$, $P < 0.0001$; θ_{EA}, $R^2 = 0.05$, $F = 8.4$, $P = 0.0042$). (There was no difference in significance for any of the tests when using nonparametric measures such as Spearman's ρ [results not shown].) To control for this correlation in looking for an association between the Tajima's D statistic and recombination, we carried out a multiple regression with both recombination rate and θ as effects for both populations. We use θ because it contains as much information about mutation as π but no information about the frequency of mutations (and, thus, we retain more statistical power).

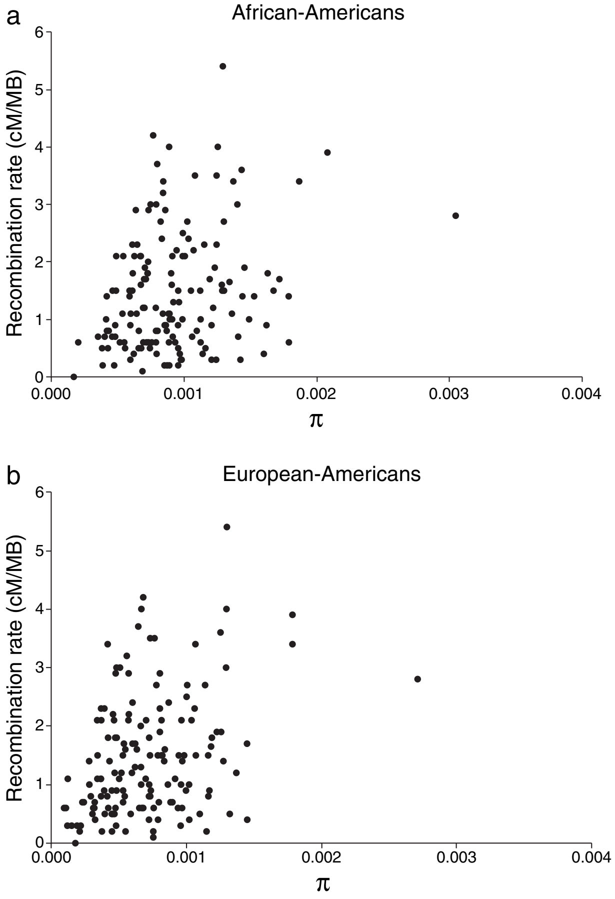

FIG. 3.—Recombination rate and level of polymorphism for all 151 loci. (*a*) African-American population; (*b*) European-American population.

There is a highly significant, positive relationship between Tajima's D statistic from loci in the European-American data set and recombination rates across loci, even while controlling for θ ($R^2 = 0.09$, $F = 14.7$, $P = 0.0002$). This relationship is not significant in the African-American data ($R^2 = 0.02$, $F = 3.1$, $P = 0.08$). If we do not control for levels of polymorphism and use simply the Tajima's D statistic values alone, there is a marginally significant relationship between recombination and D statistic values in both populations (fig. 4) (EA, $R = 0.10$, $F = 16.5$, $P < 0.0001$; AA, $R^2 = 0.03$, $F = 4.9$, $P = 0.027$). The positive relationship between the D statistic values and recombination in the European-American population suggests that multiple hitchhiking events have been associated with the migration out of Africa and colonization of novel habitats. Repeated fixation of advantageous mutations throughout the genome has caused a skew in linked variation towards lower frequencies, with more pronounced effects in regions of low recombination. The lack of a relationship in the African-American sample could be an effect of sampling from an admixed population, but the rank-order of D statistic values in this population is significantly correlated with the values in the European-American sample (Spearman's $\rho = 0.33$, $P < 0.0001$). We believe that the significant correlation present only in the European-American sample is caused by the increased number and/or effect of advantageous alleles

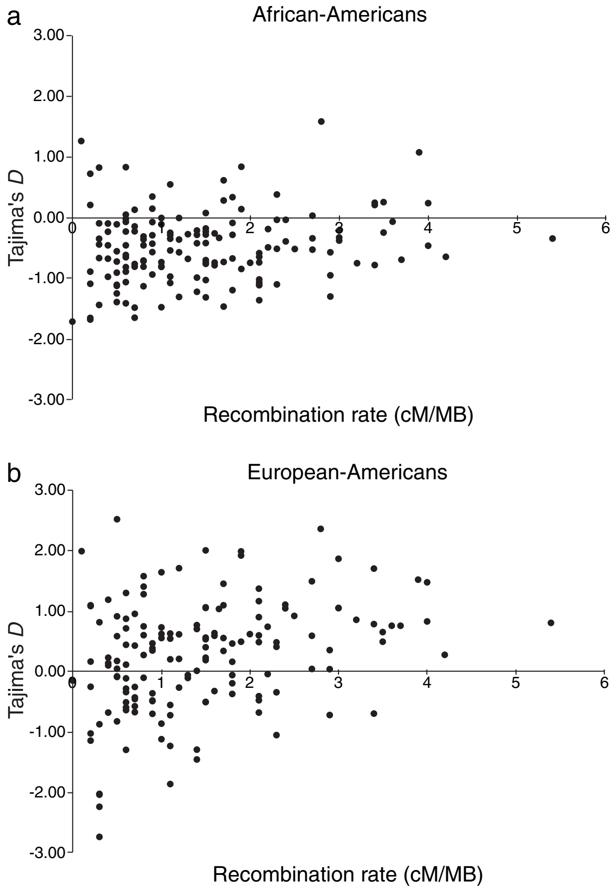

FIG. 4.—Recombination rate and Tajima's *D* statistic for all 151 loci. (*a*) African-American population; (*b*) European-American population.

that are associated with migration into new habitats. This conclusion assumes that the environments in Africa are more similar to the ancestral human environments than those found outside of Africa, and that migration out of Africa brought humans into novel environments. These conclusions are consistent, however, with a number of recent studies also reporting a preponderance of evidence for selective sweeps in non-African human populations (Kayser, Brauer, and Stoneking 2003; Storz, Payseur, and Nachman 2004). Also, because we have controlled for levels of polymorphism, the relationship between recombination and the frequency spectrum in European-Americans should be independent of any mutagenic effect of recombination (Hellmann et al. 2003).

One alternative possibility for the observed pattern is that a nonequilibrium demographic history may have caused the correlation between recombination and the frequency spectrum. To test for this effect, we generated 10,000 coalescent genealogies conditioned on the number of segregating sites observed in European-Americans and recombination rate for each of the 151 loci under our bottleneck population history, as above. We then looked at the distribution of the 10,000 Spearman's ρ values between recombination rate and Tajima's *D* statistic simulated at each locus. The correlation we observe in the European-American data is higher than all but four simulated values ($P = 0.0004$). We, therefore, have evidence that nonequilibrium processes are likely not responsible for the patterns we see.

One further alternative explanation for the correlation between recombination and Tajima's *D* statistic is background selection (Charlesworth, Morgan, and Charlesworth 1993). A significant skew in the mutation frequency spectrum is unlikely to be found with realistic sample sizes under background selection (Charlesworth, Charlesworth, and Morgan 1995), however, even with population sizes as small as are found in humans. Muller's Ratchet may also lead to a skewed frequency spectrum when the effective population size is small (Gordo, Navarro, and Charlesworth 2002), but this effect is removed by even the low levels of recombination we observe at almost all loci. We conclude, therefore, that adaptive natural selection in regions of low recombination is the most likely explanation for the observed correlations.

Hellmann et al. (2003) found no evidence for hitchhiking, as both polymorphism and divergence were correlated with recombination. Our finding of a correlation between the frequency spectrum and recombination, however, implies that there is an effect of natural selection. It appears that a combination of mutation and selection is needed to explain the observed relationships. Both the mutagenic effects of recombination and linked selection are at least partly responsible for the correlation between recombination rates and levels of polymorphism. Our results differ from similar analyses carried out in *Drosophila melanogaster*. Studies of the relationship between recombination and the mutation frequency spectrum in African and non-African populations of *D. melanogaster* have found significant correlations only in the ancestral African populations, although there is a correlation with levels of polymorphism in both populations (Aquadro, Begun, and Kindahl 1994; Andolfatto and Przeworski 2001). The lack of a significant correlation with the frequency spectrum in non-African populations has been thought to be caused by the loss of low frequency mutations in the parallel migration of *D. melanogaster* out of Africa (Andolfatto and Przeworski 2001), but we do not have any biological explanation for the apparent differences in strength or effect of the bottleneck in humans relative to flies.

The correlation between recombination and the mutation frequency spectrum highlights a caveat of demographic model fitting: genes from regions of different recombination rates will give different stories because many of the loci are affected by linked selection. On average, genes from regions with recombination rates less than 1 cM/MB ($n = 66$) have a *D* statistic value 0.85 lower in the European-American population and 0.38 lower in the African-American population than genes chosen from regions with recombination rates greater than 3 cM/MB ($n = 16$). The set of genes chosen and their local recombination rates may, therefore, determine the exact history of the sample inferred for any particular study.

Selection on ABO and TRPV6

In addition to general patterns of natural selection, we found evidence for selection on specific loci. We present evidence for the two strongest cases: one that acts as a positive control in our data set because it is a known candidate for balancing selection, and the other that we believe is a novel finding of directional positive selection. These two loci represent the opposite extremes of the distributions for Tajima's *D* statistic and, as we shall demonstrate, show significant deviations from neutrality even under severe nonequilibrium demographic scenarios. The *ABO* locus has long been thought to be under balancing selection in humans (e.g., Chung and Morton 1961; Bodmer and Cavalli-Sforza 1976; Cavalli-Sforza, Menozzi, and Piazza 1994). The presence of the A, B, and O alleles in all major human populations and low levels of differentiation among populations at this locus led researchers to hypothesize that balancing selection must be maintaining variation. This selection was variously thought to be the result of protection from disease, maternal-fetal incompatibilities, or avoidance of pathogenic infection (Chung and Morton 1961; Bodmer and Cavalli-Sforza 1976; Livingstone 1978). In a nonrandom sample of primate *ABO* sequences, Saitou and Yamamoto (1997) found an unusually high amount of nucleotide diversity, but there has never been a proper population genetic study of this locus. We found that the 22.4 kb sequenced at the *ABO* locus contained the highest levels of polymorphism in either the African-American ($\theta = 0.0020$, $\pi = 0.0030$) or European-American ($\theta = 0.0017$, $\pi = 0.0027$) populations. The value of Tajima's *D* statistic at *ABO* is also higher than that for any other locus in the African-American population ($D = +1.58$) and is higher than all but one locus (*IL1A*) in the European-American population ($D = +2.35$). The observed *D* statistic values deviate significantly from neutral expectations even when the nonequilibrium demographic models outlined above are used to generate a null coalescent distribution (AA, $P < 0.0001$; EA, $P = 0.0005$). The observed significance of the data is not strongly dependent on the precise demographic scenario we have chosen to model. If we also calculate the *P* values under an extreme, contraction-only demographic history—which will tend to skew *D* statistic values very positively—the data are still strongly significant (AA, $P = 0.0001$; EA, $P = 0.018$). The difference at *ABO* between Tajima's *D* statistic in the European-American and African-American populations (0.77) is likely caused solely by demographic effects, as it is almost exactly the average difference observed between all loci in these two populations (average $D_{EA} - D_{AA} = 0.75$). We believe that the polymorphism and frequency spectrum results both confirm the hypothesis that variation at *ABO* is maintained by balancing selection, although the exact selective mechanism remains unclear.

Unlike *ABO*, which appears to be under selection in both populations studied here, the *TRPV6* locus shows a population-specific pattern of positive selection. Where the value of Tajima's *D* statistic at *TRPV6* in the African-American population (+0.82) is not significantly different from the neutral expectation under an admixed demographic scenario, the European-American value (−2.74) is both the lowest value of any locus in this population and is highly significant under our bottleneck scenario ($P < 0.0001$), even after correction for multiple comparisons ($P < 0.01$). Again, the significance of this data is not dependent on any specific population history. An extreme model of recent population expansion with no preceding contraction—which will skew *D* statistic values very negatively—still strongly rejects the null ($P = 0.0001$). As with the *ABO* locus, most loci in our study have an increased value of Tajima's *D* statistic in the European-American population. Only 21 loci show a decrease in *D* statistic values in the European-American population relative to the African-American population. The difference in *D* statistic values between the two populations at *TRPV6* (−3.56), however, is by far the largest decrease: the next largest decrease is −1.8. We tested the significance of this difference by generating samples from 10,000 pairs of coalescent genealogies drawn from two populations simulated as in our demographic model above (one from population P_1 and one an admixture of populations P_0 and P_1) but with θ taken from *TRPV6* in the African-American sample. We then compared the differences in *D* statistic values between these paired samples. The difference observed at *TRPV6* is more negative than any of the 10,000 simulated values.

An additional signature expected by directional selection is a deficiency of haplotypes (Hudson et al. 1994; Fu 1997; Depaulis and Veuille 1998). We, therefore, used the program Phase version 1.0.1 (Stephens, Smith, and Donnelly 2001) to construct haplotypes from the unphased genotype data at this locus. Although there is some uncertainty involved in the inference of haplotypes, recent studies have shown that Phase is accurate and reasonably precise in its estimates of haplotype frequencies (Lin et al. 2002; Xu et al. 2002; Crawford et al. 2004). The *TRPV6* sample contained 29 haplotypes identical across the 29.6 kb sequenced in 46 European-American chromosomes. Using Hudson's haplotype test (Hudson et al. 1994), we asked the probability of seeing 29 identical haplotypes within a sample that contained a total of 63 segregating sites. With no recombination (a conservative assumption) we still found a highly significant deviation from the distribution of haplotypes expected under the neutral-equilibrium hypothesis ($P < 0.0001$) or after a bottleneck as modeled above ($P = 0.0009$). As expected, a population contraction and subsequent expansion had the effect of increasing linkage disequilibrium in simulated samples (Kruglyak 1999; Reich et al. 2001) but to an extent that still cannot explain the patterns at *TRPV6*.

The selective scenario we propose for the *TRPV6* locus is that a preexisting mutation in the ancestral African population became advantageous in a new environment and rose to high frequency. We favor this scenario over one involving a new mutation in European populations because there are no mutations unique to the European-American sample that have reached frequencies greater than 7%. To further investigate the origin of this inferred advantageous mutation and its associated haplotype, we constructed haplotypes at *TRPV6* using only polymorphisms shared between the European-American and African-American samples (52 segregating sites). Using only the shared polymorphisms, the European-American sample contained 42 identical haplotypes at this locus (Hudson's haplotype test for a bottlenecked population,

$P < 0.0001$). Of the four remaining haplotypes in the sample, one is different from the high-frequency haplotype at 51 of the 52 polymorphisms, including all three nonsynonymous polymorphisms (assignment probabilities given by Phase of the low-frequency mutations are all 1.00, except one mutation that has probability 0.75). It is these low-frequency mutations that are largely responsible for the extremely negative Tajima's *D* statistic value at *TRPV6*. Phasing the haplotypes at *TRPV6* in the African-American sample again using only shared polymorphisms reveals the origin of the two highly diverged haplotypic classes: both appear among African-Americans at approximately equal frequency (although there are also clear recombinant haplotypes). Therefore, we infer that *TRPV6* had two major haplotypes in Africa and that concomitant with or subsequent to migration out of Africa, a mutation associated with one of the haplotypes was driven to high frequency. This almost-complete selective sweep dragged linked mutations with it and has nearly extinguished the other haplotype. It is the remaining polymorphism on the rare haplotype, and not new mutations expected during the recovery phase from a selective sweep (Kaplan, Hudson, and Langley 1989; Braverman et al. 1995), that is largely responsible for the extremely negative *D* statistic values. It is not clear whether this combination of an almost-completed sweep with a skew in the frequency spectrum is to be generally expected (Fu 1997). The preponderance of evidence from multiple analyses strongly suggests that *TRPV6*, or a gene in linkage disequilibrium with it, is under positive directional selection in the European-American population.

Independence of Loci

The question of whether *TRPV6* or some other linked locus is the actual target of selection raises an important issue in the analysis of population genomic data sets: the nonindependence of loci. Physical linkage and linkage disequilibrium among loci means that assumptions about independence do not hold for large data sets. Calculating values of Tajima's *D* statistic (or many other statistics) for 30,000 genes will almost certainly uncover correlations among linked loci caused by shared histories. Fitting neutral demographic models to data that are not independent, much less neutral, will also be a problem with large data sets (Wall and Przeworski 2000). Even within the data set analyzed here, there are 19 chromosomes, 28 chromosome arms, and 32 chromosome bands that contain multiple genes. Although these are somewhat arbitrary units of clustering, there is evidence for significant effects of each.

We can ask whether there are chromosome-specific, arm-specific, or band-specific measures of polymorphism in either of the two populations or in differences between the populations. Using only the chromosomes and arms represented by multiple genes, there are significant effects in analyses of variance of each on some of the measures of polymorphism (all $P < 0.05$; chromosome: π_{AA}, $\theta_{AA} - \theta_{EA}$, and $\pi_{AA} - \pi_{EA}$; arm: π_{AA}, $\theta_{AA} - \theta_{EA}$, $\pi_{AA} - \pi_{EA}$, and D_{AA}). If we use only the chromosome bands that contain multiple genes, there are significant effects on π_{EA}, $\pi_{AA} - \pi_{EA}$, $\theta_{AA} - \theta_{EA}$, D_{EA}, and $D_{EA} - D_{AA}$. The presence of effects of band on both levels of polymorphism and Tajima's *D* statistic suggests that these effects are caused by selective events in the European-American population. Alternatively, they could simply be a result of sampling variance among shorter stretches of DNA during migration out of Africa.

We believe that selection almost certainly explains some of the effect of chromosome bands. The band that contains the *TRPV6* locus, 7q34, shows a significantly lower Tajima's *D* statistic value in the European-American population than any other band (Student's *t*-test, $P < 0.05$). The *KEL, EPHB6*, and *TRPV5* genes are also in this band and also show highly skewed mutation frequency spectra in European-Americans ($D = -2.24$, -2.03, and -2.04, respectively). These four genes are all contained within a 100-kb stretch of DNA that has a relatively low recombination rate (0.3 cM/MB). There is significant linkage disequilibrium between loci: we were able to construct strongly supported haplotypes across this region. Using only those polymorphisms shared between the two populations for all four loci, we found a single 100-kb haplotype that was identical in 14 of the 46 European-American chromosomes (there were 160 total polymorphisms in this reduced sample). This high-frequency haplotype is highly unlikely, even after a bottleneck and subsequent recovery (Hudson's haplotype test, $P = 0.002$). Although *TRPV6* shows the most extreme values for the mutation and haplotype frequency spectra and for the reduction in *D* statistic value in European-Americans relative to African-Americans (*KEL* $= -1.8$, *EPHB6* $= -1.36$, and *TRPV5* $= -1.69$), linkage disequilibrium across the region implies that it may be another gene that is the true target of positive selection. The *KEL* blood group locus is an interesting candidate as a target of selection but shows no amino acid polymorphisms that are associated with the high-frequency haplotype. There may also be genes outside the 100-kb region studied here that are the target of natural selection; the size of the region affected remains to be established. Because of the extreme heterogeneity in recombination rates along human chromosomes, linkage disequilibrium around selected genes has been found to extend for hundreds of thousands of bases (e.g., Tishkoff et al. 1996; Sabeti et al. 2002; Saunders, Hammer, and Nachman 2002). If selective events in humans are common and of large effect, it may be that a considerable proportion of the variation in the human genome has been shaped by linked selection.

An important caveat follows from these results: nonindependence among loci may have violated some of the assumptions of our analyses. Sampling of linked loci will act to pseudoreplicate relationships present in only a fraction of the genome. Fortunately, the evidence for pseudoreplication as being a factor in our analyses is relatively weak. Only 52 of the 151 loci reside in a chromosome band by themselves, but other than band 7q34, there are no bands that are individually different from the rest for any statistic. Also, the relationship that we found between Tajima's *D* statistic values and recombination while controlling for θ remains significant after removing the four loci within band 7q34 ($R^2 = 0.06$,

$F = 9.25$, and $P = 0.0028$). Nonindependence among loci may also have had a conservative effect on our conclusions: if tests on individual loci are not truly independent, then correcting for 151 independent comparisons is a very conservative approach. We only had strong statistical support for the two loci at the ends of the distribution of *D* statistic values, *ABO* and *TRPV6*. The necessity of correcting for multiple comparisons will always be a pitfall in population genomics, but simply looking in the tails of the distribution may also incorrectly identify targets of natural selection. The second, third, and fourth lowest values of Tajima's *D* statistic in European-Americans are all in significant linkage disequilibrium with *TRPV6*. It should also be pointed out that the complications of nonindependence will be less acute when dealing with tests or statistics that compare classes of mutations within a locus (e.g., McDonald and Kreitman 1991). This is because these measures will not be strongly affected by linkage disequilibrium with linked loci (McDonald and Kreitman 1991). For measures that do not have this advantage, however, it may be that spatially explicit analyses across chromosomes will be necessary.

Conclusions

Whereas demographic events such as population bottlenecks or expansions will affect all genes in a genome, natural selection is expected to have only locus-specific or region-specific effects on DNA variation. Our analyses have shown that the demographic histories of human populations can largely account for the level and frequency of variation across the genome. However, even working within a nonequilibrium framework, we were able to show deviations from neutral expectations at the *ABO* and *TRPV6* loci and in many regions of low recombination. The results for this data set are consistent with the combined effects of a population bottleneck and repeated selective sweeps in the human migration out of Africa—in agreement with previous reports (Kayser, Brauer, and Stoneking 2003; Storz, Payseur, and Nachman 2004)—and suggest that natural selection affects a relatively large proportion of the genome.

Acknowledgments

We gratefully acknowledge D. Begun and C. Langley for many comments and discussion of the ideas presented here. We also thank J. Gillespie, C. Jones, A. Kern, L. Moyle, M. Sanderson, J. Wares, and J. Wilkins for insightful suggestions and R. Hudson, J. Mezey, and K. Thornton for sharing their code. D. Goldstein and an anonymous reviewer offered comments that substantially improved the manuscript. We acknowledge support from a National Science Foundation predoctoral fellowship (J.E.S.) and from a National Science Foundation Integrative Informatics postdoctoral fellowship (M.W.H.).

Literature Cited

Akey, J. M., G. Zhang, K. Zhang, L. Jin, and M. D. Shriver. 2002. Interrogating a high-density SNP map for signatures of natural selection. Genome Res. **12**:1805–1814.

Andolfatto, P., and M. Przeworski. 2001. Regions of lower crossing over harbor more rare variants in African populations of *Drosophila melanogaster*. Genetics **158**:657–665.

Aquadro, C. F., D. J. Begun, and E. C. Kindahl. 1994. Selection, recombination, and DNA polymorphism in *Drosophila*. Pp. 46–56 *in* B. Golding, ed. Non-neutral evolution. Chapman and Hall, New York.

Begun, D. J., and C. F. Aquadro. 1992. Levels of naturally occurring DNA polymorphism correlate with recombination rates in *Drosophila melanogaster*. Nature **356**:519–520.

Begun, D. J., and P. Whitley. 2000. Reduced X-linked nucleotide polymorphism in *Drosophila simulans*. Proc. Natl. Acad. Sci. USA **97**:5960–5965.

Birky, C. W. Jr, and J. B. Walsh. 1988. Effects of linkage on rates of molecular evolution. Proc. Natl. Acad. Sci. USA **85**: 6414–6418.

Bodmer, W. F., and L. L. Cavalli-Sforza. 1976. Genetics, evolution, and man. W.H. Freeman and Company, San Francisco.

Bowcock, A. M., J. R. Kidd, J. L. Mountain, J. M. Hebert, L. Carotenuto, K. K. Kidd, and L. L. Cavalli-Sforza. 1991. Drift, admixture, and selection in human evolution: a study with DNA polymorphisms. Proc. Natl. Acad. Sci. USA **88**: 839–843.

Braverman, J. M., R. R. Hudson, N. L. Kaplan, C. H. Langley, and W. Stephan. 1995. The hitchhiking effect on the site frequency spectrum of DNA polymorphisms. Genetics **140**:783–796.

Cann, R. L., M. Stoneking, and A. C. Wilson. 1987. Mitochondrial DNA and human evolution. Nature **325**:31–36.

Cargill, M., D. Altshuler, J. Ireland et al. (17 co-authors). 1999. Characterization of single-nucleotide polymorphisms in coding regions of human genes. Nat. Genet. **22**:231–238.

Carlson, C. S., M. A. Eberle, M. J. Rieder, J. D. Smith, L. Kruglyak, and D. A. Nickerson. 2003. Additional SNPs and linkage-disequilibrium analyses are necessary for whole-genome association studies in humans. Nat. Genet. **33**:518–521.

Carlson, C. S., M. A. Eberle, M. J. Rieder, Q. Yi, L. Kruglyak, and D. A. Nickerson. 2004. Selecting a maximally informative set of single-nucleotide polymorphisms for association analyses using linkage disequilibrium. Am J. Hum. Genet. **74**:106–120.

Cavalli-Sforza, L. L., P. Menozzi, and A. Piazza. 1994. The history and geography of human genes. Princeton University Press, Princeton, New Jersey.

Chakraborty, R., M. I. Kamboh, M. Nwankwo, and R. E. Ferrell. 1992. Caucasian genes in American blacks: new data. Am. J. Hum. Genet. **50**:145–155.

Charlesworth, B., M. T. Morgan, and D. Charlesworth. 1993. The effect of deleterious mutations on neutral molecular variation. Genetics **134**:1289–1303.

Charlesworth, D., B. Charlesworth, and M. T. Morgan. 1995. The pattern of neutral molecular variation under the background selection model. Genetics **141**:1619–1632.

Chung, C. S., and N. E. Morton. 1961. Selection at the ABO locus. Am. J. Hum. Genet. **13**:9–27.

Crawford, D. C., C. S. Carlson, M. J. Rieder, D. P. Carrington, Q. Yi, J. D. Smith, M. A. Eberle, L. Kruglyak, and D. A. Nickerson. 2004. Haplotype diversity across 100 candidate genes for inflammation, lipid metabolism, and blood pressure regulation in two populations. Am. J. Hum. Genet. **74**: 610–622.

Depaulis, F., and M. Veuille. 1998. Neutrality tests based on the distribution of haplotypes under an infinite-site model. Mol. Biol. Evol. **15**:1788–1790.

Dvorak, J., M. C. Luo, and Z. L. Yang. 1998. Restriction fragment length polymorphism and divergence in the genomic regions of high and low recombination in self-fertilizing and cross-fertilizing *Aegilops* species. Genetics **148**:423–434.

Excoffier, L., and S. Schneider. 1999. Why hunter-gatherer populations do not show signs of Pleistocene demographic expansions. Proc. Natl. Acad. Sci. USA **96**:10597–10602.

Fay, J. C., and C. I. Wu. 1999. A human population bottleneck can account for the discordance between patterns of mitochondrial versus nuclear DNA variation. Mol. Biol. Evol. **16**:1003–1005.

Frisse, L., R. R. Hudson, A. Bartoszewicz, J. D. Wall, J. Donfack, and A. Di Rienzo. 2001. Gene conversion and different population histories may explain the contrast between polymorphism and linkage disequilibrium levels. Am. J. Hum. Genet. **69**:831–843.

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

Fu, Y. X. 1997. Statistical tests of neutrality of mutations against population growth, hitchhiking and background selection. Genetics **147**:915–925.

Galtier, N., F. Depaulis, and N. H. Barton. 2000. Detecting bottlenecks and selective sweeps from DNA sequence polymorphism. Genetics **155**:981–987.

Gillespie, J. H. 1991. The causes of molecular evolution. Oxford University Press, New York.

———. 2000. Genetic drift in an infinite population: the pseudohitchhiking model. Genetics **155**:909–919.

Glinka, S., L. Ometto, S. Mousset, W. Stephan, and D. De Lorenzo. 2003. Demography and nautral selection have shaped genetic variation in *Drosophila melanogaster*: a multilocus approach. Genetics **165**:1269–1278.

Goldstein, D. B., and L. Chikhi. 2002. Human migrations and population structure: what we know and why it matters. Ann. Rev. Genomics Hum. Genet. **3**:129–152.

Gordo, I., A. Navarro, and B. Charlesworth. 2002. Muller's ratchet and the pattern of variation at a neutral locus. Genetics **161**:835–848.

Hahn, M. W., M. D. Rausher, and C. W. Cunningham. 2002. Distinguishing between selection and population expansion in an experimental lineage of bacteriophage T7. Genetics **161**:11–20.

Hahn, M. W., M. V. Rockman, N. Soranzo, D. B. Goldstein, and G. A. Wray. 2004. Population genetic and phylogenetic evidence for positive selection on regulatory mutations at the *Factor VII* locus in humans. Genetics **167**:867–877.

Halushka, M. K., J. B. Fan, K. Bentley, L. Hsie, N. P. Shen, A. Weder, R. Cooper, R. Lipshutz, and A. Chakravarti. 1999. Patterns of single-nucleotide polymorphisms in candidate genes for blood-pressure homeostasis. Nat. Genet. **22**: 239–247.

Harpending, H., and A. Rogers. 2000. Genetic perspectives on human origins and differentiation. Annu. Rev. Genomics Hum. Genet. **1**:361–385.

Harpending, H. C., M. A. Batzer, M. Gurven, L. B. Jorde, A. R. Rogers, and S. T. Sherry. 1998. Genetic traces of ancient demography. Proc. Natl. Acad. Sci. USA **95**: 1961–1967.

Hellmann, I., I. Ebersberger, S. E. Ptak, S. Paabo, and M. Przeworski. 2003. A neutral explanation for the correlation of diversity with recombination rates in humans. Am. J. Hum. Genet. **72**:1527–1535.

Hudson, R. R. 1990. Gene genealogies and the coalescent process. Pp. 1–44 *in* D. Futuyma and J. Antonovics, eds. Oxford surveys in evolutionary biology. Oxford University Press, Oxford, UK.

———. 2002. Generating samples under a Wright-Fisher neutral model of genetic variation. Bioinformatics **18**:337–338.

Hudson, R. R., K. Bailey, D. Skarecky, J. Kwiatowski, and F. J. Ayala. 1994. Evidence for positive selection in the superoxide dismutase (*Sod*) regions of *Drosophila melanogaster*. Genetics **136**:1329–1340.

Hudson, R. R., and N. L. Kaplan. 1994. Gene trees with background selection. Pp. 140–153 *in* B. Golding, ed. Nonneutral evolution. Chapman and Hall, New York.

Jorde, L. B., A. R. Rogers, M. Bamshad, W. S. Watkins, P. Krakowiak, S. Sung, J. Kere, and H. C. Harpending. 1997. Microsatellite diversity and the demographic history of modern humans. Proc.Natl. Acad. Sci. USA **94**:3100–3103.

Kaplan, N. L., R. R. Hudson, and C. H. Langley. 1989. The "hitchiking effect" revisited. Genetics **123**:887–899.

Kayser, M., S. Brauer, and M. Stoneking. 2003. A genome scan to detect candidate regions influenced by local natural selection in human populations. Mol. Biol. Evol. **20**:893–900.

Kimmel, M., R. Chakraborty, J. P. King, M. Bamshad, W. S. Watkins, and L. B. Jorde. 1998. Signatures of population expansion in microsatellite repeat data. Genetics **148**: 1921–1930.

Kong, A., D. F. Gudbjartsson, J. Sainz et al. (16 co-authors). 2002. A high-resolution recombination map of the human genome. Nat. Genet. **31**:241–247.

Kraft, T., T. Sall, I. Magnusson-Rading, N. O. Nilsson, and C. Hallden. 1998. Positive correlation between recombination rates and levels of genetic variation in natural populations of sea beet (*Beta vulgaris* subsp. *maritima*). Genetics **150**: 1239–1244.

Kruglyak, L. 1999. Prospects for whole-genome linkage disequilibrium mapping of common disease genes. Nat. Genet. **22**:139–144.

Lin, S., D. J. Cutler, M. E. Zwick, and A. Chakravarti. 2002. Haplotype inference in random population samples. Am. J. Hum. Genet. **71**:1129–1137.

Livingstone, F. B. 1978. Frequency-dependent selection and the ABO blood groups. Pp. 127–139 *in* R. J. Meier, C. M. Otten, and F. Abdel-Hameed, eds. Evolutionary models and studies in human diversity. Mouton Publishers, Paris.

Marth, G., G. Schuler, R. Yeh et al. (20 co-authors). 2003. Sequence variations in the public human genome data reflect a bottlenecked population history. Proc Natl. Acad. Sci. USA **100**:376–381.

Marth, G. T., E. Czabarka, J. Murvai, and S. T. Sherry. 2004. The allele frequency spectrum in genome-wide human variation data reveals signals of differential demographic history in three large world populations. Genetics **166**:351–372.

Maynard Smith, J., and J. Haigh. 1974. The hitch-hiking effect of a favorable gene. Genet. Res. **23**:23–35.

McDonald, J. H., and M. Kreitman. 1991. Adaptive protein evolution at the Adh locus in *Drosophila*. Nature: 652–654.

Nachman, M. W. 1997. Patterns of DNA variability at X-linked loci in *Mus domesticus*. Genetics **147**:1303–1316.

———. 2001. Single nucleotide polymorphisms and recombination rate in humans. Trends Genet. **17**:481–485.

Parra, E. J., A. Marcini, J. Akey et al. (11 co-authors). 1998. Estimating African American admixture proportions by use of population-specific alleles. Am. J. Hum. Genet. **63**: 1839–1851.

Payseur, B. A., and M. W. Nachman. 2002. Natural selection at linked sites in humans. Gene **300**:31–42.

Pluzhnikov, A., A. Di Rienzo, and R. R. Hudson. 2002. Inferences about human demography based on multilocus analyses of noncoding sequences. Genetics **161**:1209–1218.

Przeworski, M., R. R. Hudson, and A. Di rienzo. 2000. Adjusting the focus on human variation. Trends Genet. **16**:296–302.

Przeworksi, M., J. D. Wall, and P. Andolfatto. 2001. Recombination and the frequency spectrum in *Drosophila melanogaster* and *Drosophila simulans*. Mol. Biol. Evol. **18**:291–298.

Ptak, S. E., and M. Przeworski. 2002. Evidence for population growth in humans is confounded by fine-scale population structure. Trends Genet. **18**:559–563.

Reich, D. E., M. Cargill, S. Bolk et al. (11 co-authors). 2001. Linkage disequilibrium in the human genome. Nature **411**:199–204.

Rockman, M. V., M. W. Hahn, N. Soranzo, D. B. Goldstein, and G. A. Wray. 2003. Positive selection on a human-specific transcription factor binding site regulating *IL4* expression. Curr. Biol. **13**:2118–2123.

Sabeti, P. C., D. E. Reich, J. M. Higgins et al. (17 co-authors). 2002. Detecting recent positive selection in the human genome from haplotype structure. Nature **419**:832–837.

Sachidanandam, R., D. Weissman, S. C. Schmidt et al. (41 co-authors). 2001. A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. Nature **409**:928–933.

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

Saunders, M. A., M. F. Hammer, and M. W. Nachman. 2002. Nucleotide variability at *G6pd* and the signature of malarial selection in humans. Genetics **162**:1849–1861.

Shen, P., M. Buchholz, R. Sung, A. Roxas, C. Franco, W.-H. Yang, R. Jagadeesan, K. Davis, and P. J. Oefner. 2002. Population genetic implications from DNA polymorphism in random human genomic sequences. Hum. Mutat. **20**:209–217.

Simonsen, K. L., G. A. Churchill, and C. F. Aquadro. 1995. Properties of statistical tests of neutrality for DNA polymorphism data. Genetics **141**:413–429.

Sokal, R. R., and F. J. Rohlf. 1995. Biometry. W.H. Freeman, New York.

Stajich, J. E., D. Block, K. Boulez et al. (21 co-authors). 2002. The bioperl toolkit: Perl modules for the life sciences. Genome Res. **12**:1611–1618.

Stephan, W., and C. H. Langley. 1989. Molecular genetic variation in the centromeric region of the X chromosome in three *Drosophila ananassae* populations. I. Contrasts between the *vermillion* and *forked* loci. Genetics **121**:89–99.

———. 1998. DNA polymorphism in *Lycopersicon* and crossing-over per physical length. Genetics **150**:1585–1593.

Stephens, M., N. J. Smith, and P. Donnelly. 2001. A new statistical method for haplotype reconstruction from population data. Am. J. Hum. Genet. **68**:978–989.

Storz, J. F., B. A. Payseur, and M. W. Nachman. 2004. Genome scans of DNA variability in humans reveal evidence for selective sweeps outside of Africa. Mol. Biol. Evol. **21**:1800–1811.

Tajima, F. 1983. Evolutionary relationship of DNA sequences in finite populations. Genetics **105**:437–460.

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

Tenaillon, M. I., M. C. Sawkins, A. D. Long, R. L. Gaut, J. F. Doebley, and B. S. Gaut. 2001. Patterns of DNA sequence polymorphism along chromosome 1 of maize (*Zea mays* ssp *mays* L.). Proc. Natl. Acad. Sci. USA **98**:9161–9166.

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

Tishkoff, S. A., E. Dietzsch, W. Speed et al. (15 co-authors). 1996. Global patterns of linkage disequilibrium at the CD4 locus and modern human origins. Science **271**:1380–1387.

Wakeley, J. 2000. The effects of subdivision on the genetic divergence of populations and species. Evolution **54**:1092–1101.

Wall, J. D., P. Andolfatto, and M. Przeworski. 2002. Testing models of selection and demography in *Drosophila simulans*. Genetics **162**:203–216.

Wall, J. D., and M. Przeworski. 2000. When did the human population size start increasing? Genetics **155**:1865–1874.

Watterson, G. A. 1975. On the number of segregating sites in genetical models without recombination. Theor. Pop. Biol. **7**:256–275.

Weir, B. S. 1996. Genetic data analysis II. Sinauer Associates, Sunderland, Mass.

Wright, S. 1951. The genetical structure of populations. Ann. Eugen. **15**:323–354.

Xu, C. F., K. Lewis, K. L. Cantone, P. Khan, C. Donnelly, N. White, N. Crocker, P. R. Boyd, D. V. Zaykin, and I. J. Purvis. 2002. Effectiveness of computational methods in haplotype prediction. Hum. Genet. **110**:148–156.

David Goldstein, Associate Editor

Accepted August 31, 2004