DNA Variability and Recombination Rates at *X*-Linked Loci in Humans

Michael W. Nachman,* Vanessa L. Bauer,† Susan L. Crowell* and Charles F. Aquadro†

**Department of Ecology and Evolutionary Biology, University of Arizona, Tucson, Arizona 85721 and †Section of Genetics and Development, Cornell University, Ithaca, New York 14853*

Manuscript received March 30, 1998
Accepted for publication July 20, 1998

ABSTRACT

We sequenced 11,365 bp from introns of seven *X*-linked genes in 10 humans, one chimpanzee, and one orangutan to (i) provide an average estimate of nucleotide diversity (π) in humans, (ii) investigate whether there is variation in π among loci, (iii) compare ratios of polymorphism to divergence among loci, and (iv) provide a preliminary test of the hypothesis that heterozygosity is positively correlated with the local rate of recombination. The average value for π was low (0.063%, SE = 0.036%), about one order of magnitude smaller than for *Drosophila melanogaster*, the species for which the best data are available. Among loci, π varied by over one order of magnitude. Statistical tests of neutrality based on ratios of polymorphism to divergence or based on the frequency spectrum of variation within humans failed to reject a neutral, equilibrium model. However, there was a positive correlation between heterozygosity and rate of recombination, suggesting that the joint effects of selection and linkage are important in shaping patterns of nucleotide variation in humans.

AN accurate description of the level and pattern of genetic variation in natural populations is a prerequisite to understanding the genetic basis of evolution. Kreitman (1983) published the first description of DNA sequence variation from a sample of alleles taken from a natural population. Since then, contributions from a number of laboratories have helped present a reasonably detailed picture of DNA sequence variation from multiple loci in natural populations of *Drosophila melanogaster* (reviewed in Aquadro 1992, 1997; Kreitman and Akashi 1995; Moriyama and Powell 1996). One important observation to emerge from this work is that there is considerable heterogeneity among genes in the level of naturally occurring DNA polymorphism. Moreover, the level of polymorphism at different loci is positively correlated with the local rate of recombination (Begun and Aquadro 1992; Aquadro *et al.* 1994). Because interspecific divergence is not correlated with recombination rate, such a pattern is inconsistent with a model of neutral molecular evolution. However, a positive correlation between polymorphism and recombination rate may be consistent with purifying selection against deleterious mutations and the removal of linked neutral variants (*i.e.*, background selection; Charlesworth *et al.* 1993, 1995; Charlesworth 1994; Hudson 1994; Hudson and Kaplan 1994, 1995) or with positive selection causing the fixation of adaptive mutations and the fixation of linked neutral variants (*i.e.*, genetic hitchhiking; Maynard Smith and Haigh 1974; Kaplan *et al.* 1989). Regardless of the relative contributions of these two processes it is clear that in *D. melanogaster*, (i) the interaction of natural selection and recombination is important in determining levels of variation at nuclear genes, and (ii) different genes have different evolutionary histories and experience different effective population sizes depending on their recombinational environment.

How much DNA sequence variation exists at nuclear genes in humans and what is the organization of that variation? Surprisingly, our understanding of the structure and evolutionary significance of human nucleotide polymorphism is still in its infancy. Li and Sadler (1991) compared sequences from the database for 49 loci where multiple sequences (usually two) were available and found that nucleotide diversity at fourfold degenerate sites (0.11%) was about one order of magnitude lower than in *D. melanogaster*. This study provided an important first glimpse at the average level of variability in humans, but because of small sample sizes it did not permit comparisons among loci nor did it provide a picture of the organization of variation at any single locus. Several studies have also described low levels of variation in the nonrecombining portion of the *Y* chromosome (Dorit *et al.* 1995; Hammer 1995; Whitfield *et al.* 1995). Hey (1997) described variation in a worldwide sample of eight individuals for 1.8 kb at *Pdha1*, an *X*-linked locus, and suggested that human nuclear genes may show more intermediate-frequency polymorphisms than mitochondrial genes, although the generality of this proposition remains to be tested. One of the most comprehensive surveys of nucleotide variation from a single locus comes from studies of the β-globin gene on chromosome *11* (Fullerton *et al.* 1994; Fullerton

Corresponding author: Michael W. Nachman, Department of Ecology and Evolutionary Biology, Biosciences West Bldg., University of Arizona, Tucson, AZ 85721. E-mail: nachman@u.arizona.edu

1996; Harding *et al.* 1997). A 3-kb stretch including the gene and flanking regions has been sequenced in 349 chromosomes from nine populations in Africa, Asia, and Europe, revealing an overall level of nucleotide diversity (0.18%) slightly higher than the average value at fourfold degenerate sites (0.11%) reported by Li and Sadler (1991). Another recent, large survey of a nuclear locus involves the lipoprotein lipase gene on chromosome *8* (Clark *et al.* 1998; Nickerson *et al.* 1998). A 9.7-kb region was sequenced in 142 chromosomes, revealing an average nucleotide diversity of 0.2%.

Here we report DNA sequence variation from introns of seven *X*-linked loci in 10 humans, one chimpanzee, and one orangutan. First, we ask whether there is significant heterogeneity among loci in the level of heterozygosity. Second, we compare levels of intra- and interspecific variation among loci to test the neutral prediction of equal ratios of polymorphism to divergence (Hudson *et al.* 1987). Third, the increased integration of genetic and physical maps in humans indicates that, as in Drosophila, rates of recombination vary dramatically across the genome. We provide a first look at whether heterogeneity in levels of nucleotide polymorphism is correlated with this recombinational variation. We find that levels of polymorphism differ by more than one order of magnitude among genes. However, there is no statistical support for different ratios of polymorphism to divergence, due in part to the overall low level of heterozygosity. The present results (with seven genes) reveal a positive correlation between heterozygosity and recombination rate, although sampling of additional loci is needed to confirm this observation.

MATERIALS AND METHODS

Samples: We focused on the *X* chromosome because it is well mapped, both physically and genetically (Wang *et al.* 1994; Roest Crollius *et al.* 1996; Nagaraja *et al.* 1997), and because single alleles could be PCR-amplified from males, thus avoiding the problems of sequencing and scoring heterozygous sites. Genomic DNAs representing 10 individuals (5 from Africa, 2 from North America, 2 from South America, and 1 from Europe; Table 1) were selected from the *Y* Chromosome Consortium (YCC) DNA repository provided by Dr. M. F. Hammer. One of these cell lines (Yale 117, Table 1) was subsequently shown to derive from a female using X- and Y-specific PCR primers (M. F. Hammer, personal communication). No heterozygous sites were identified from direct sequencing of PCR products; we analyzed our data assuming one allele was sampled from this individual. A single common chimpanzee (*Pan troglodytyes*) and a single orangutan (*Pongo pygmaeus*) were also surveyed from DNAs provided by Dr. O. A. Ryder.

Estimation of recombination rates: The genetic and physical map distances from the *Xp* telomere for 249 loci (Wang *et al.* 1994) are shown in Figure 1. The slope of this curve (ratio of genetic to physical distance) gives the rate of recombination for different regions of the *X* chromosome. Recombination rates are substantially reduced in a region surrounding the centromere (at 62.0 Mb) and near the *Xist* locus (at 81.2 Mb in Xq13.2), while several regions on both the *p* and *q* arms show elevated recombination rates (Wang *et al.* 1994; Nagaraja *et al.* 1997). We chose seven genes from regions that experience different rates of recombination for our survey of polymorphism (detailed below). Rates of recombination for each of these genes were calculated as the derivative of a third-order polynomial for a 5-cM window centered on the locus of interest. Linear models fit to these data gave similar estimates of recombination rate and did not change any of the conclusions.

PCR amplification and DNA sequencing: We surveyed introns from pyruvate dehydrogenase α-subunit (*Pdha1*), glycerol kinase (*Gk*), dystrophin (*Dmd*), interleukin-2 receptor γ chain (*Il2rg*), myelin proteolipid protein (*Plp*), hypoxanthine phosphoribosyltransferase (*Hprt*), and iduronate sulphate sulphatase (*Ids*). Introns were surveyed so as to maximize the potential for detecting differences in levels of polymorphism with minimum complications due to differing levels of functional constraint. Primers were typically placed in exons to amplify intervening introns, and in some cases, primers were placed in introns; in all cases, primers were designed to lie in conserved regions based on interspecific comparisons when available. Primer sequences were designed from published sequences of *Pdha1* (introns 9 and 10; Koike *et al.* 1990), *Gk* (intron 1; Guo *et al.* 1993), *Dmd* (intron 44; Richards 1992), *Il2rg* (introns 4 and 5; Noguchi *et al.* 1993), *Plp* (intron 5; Diehl *et al.* 1986), *Hprt* (introns 2 and 8; Edwards *et al.* 1990), and *Ids* (intron 5; Lu *et al.* 1994).

DNA was amplified using PCR (Saiki *et al.* 1988) in 100-μl reaction volumes with 40 cycles of 94° 1 min, 55° 1 min, and 72° 2 min. Taq polymerase (Perkin Elmer-Cetus, Norwalk, CT) was used with conditions as specified by the supplier, and the reaction mixture was overlaid with mineral oil. Following the reaction, oil was removed using a chloroform extraction, and the double-stranded PCR product was precipitated with 33 μl ammonium acetate (10 m) and 133 μl 100% cold ethanol, washed once in 80% ethanol, and resuspended in 30 μl ddH_2O. Double-stranded PCR products were sequenced directly using the dideoxy chain termination method (Sanger *et al.* 1977) using Sequenase (Amersham, Arlington Heights, IL) with slight modifications as described in Nachman (1997), or in cycle-sequencing reactions using Thermosequenase (Amersham) with conditions as specified by the supplier. Sequencing primers were typically 17-mers and were spaced approximately every 200 bp. A total of 11,365 bp was sequenced in each human, one chimpanzee, and one orangutan. This is an average of 1.6 kb per locus; exact numbers for each locus are given in Table 2.

Data analysis: Sequences were aligned by eye, and the numbers and frequencies of all polymorphic sites were counted. Two measures of nucleotide variability, π (Nei and Li 1979) and θ (Watterson 1975), were calculated for each locus. Nucleotide diversity, π, is based on the average number of nucleotide differences between two sequences randomly drawn from a sample, and θ is based on the proportion of segregating sites in a sample. Variances of θ and π were calculated (Watterson 1975; Tajima 1983; Hartl and Clark 1997). Under equilibrium conditions with respect to mutation and drift, both π and θ estimate the neutral parameter $3N_e\mu$ for *X*-linked loci, where N_e is the effective population size and μ is the neutral mutation rate. Tajima's *D* statistic (Tajima 1989) was calculated to test this neutral expectation for each locus. Fu and Li's *D* (1993) was also calculated to test for deviations from a neutral frequency distribution. Linkage disequilibrium (D') was calculated for a set of independent pairwise comparisons between polymorphic sites among the 10 individuals (Lewontin 1964), and the significance of D' was assessed using Fisher's exact tests. Ratios of polymorphism to divergence were compared with the expectations under a

TABLE 1

Polymorphic nucleotide sites at introns of *X*-linked genes in humans

Identity	Geographic origin	Ethnic group	*Pdhal* 00011 26935 58687 13420	*Gk* 0 0 5 1	*Dmd* 00000111 16789113 65800392 33660265	*Plp* 00 24 43 61	*Hprt* 0001 4670 4019 1771
Consensus			CCGTT	C	ACTTCAGG	TA	GAGA
JR013	Namibia	Tsumkwe	.....	.	..A....A	..	....
JR323	Namibia	Tsumkwe	T.CCC	.	.G.C.TT.	..	A.A.
Alb77	South Africa	S. Sotho	T.CCC	.	GG...T..	.C	....
Alb74	South Africa	Pedi	T.CCC	.	G.....	..	.C..
LD156	South Africa	Herero	TT.CC	.	G...T...	..	.C..
Yale117	North America	Amerindian	.....	G	.....	G.	....
TiMu/AR01	North America	Navaho	.....	G	G.....	G.	....
JK1364	Brazil	Karitiana	.....	.	.G.....	..	....
JK1370	Brazil	Karitiana	.....	.	.G.....	G.	...C
SaGr	Meditteranean	Ashkenazi	.....	.	.G.C.TT.	G.	....

Genes are listed in order of genetic and physical location on the *X* chromosome (from *p* to *q* telomeres).

neutral model using a seven-locus Hudson, Kreitman, and Aguade (HKA) test (Hudson *et al.* 1987). Coalescence times for each locus were calculated as in Fu (1996) and Fu and Li (1997). Sequences have been submitted to GenBank under accession numbers AF085430–AF085462.

RESULTS

Recombination rates, levels of polymorphism, and levels of divergence for each of the genes are summarized in Table 2. We observed a total of 20 polymorphic nucleotide sites (Table 1), and the average level of nucleotide diversity was very low ($\pi = 0.063\%$). However, nucleotide diversity varied by more than one order of magnitude among the genes, from 0 at *Il2rg* and *Ids* to a high value of 0.187% at *Dmd*. No insertion/deletion polymorphisms were observed within humans. For most loci values of π and θ were similar to each other (Table 2). Neither Tajima's *D* nor Fu and Li's *D* differed significantly from the neutral expectation of 0 for any of the loci. These data provide no support for the idea that nuclear genes in humans typically show an excess of intermediate-frequency polymorphisms (Hey 1997).

Recombination rates varied from <1 cM/Mb at *Il2rg* and *Plp* to over 7 cM/Mb at *Dmd*. Measuring the extent of linkage disequilibrium in these data is complicated by the small number of segregating sites and the small sample size. Excluding polymorphisms present in only one individual (*i.e.*, singletons), there were 12 polymorphic sites (Table 1). It is common to compute D' (Lewontin 1964) for all $n(n-1)/2$ pairwise comparisons among n segregating sites in a sample to look for groups of sites in linkage disequilibrium (*e.g.*, Schaeffer and

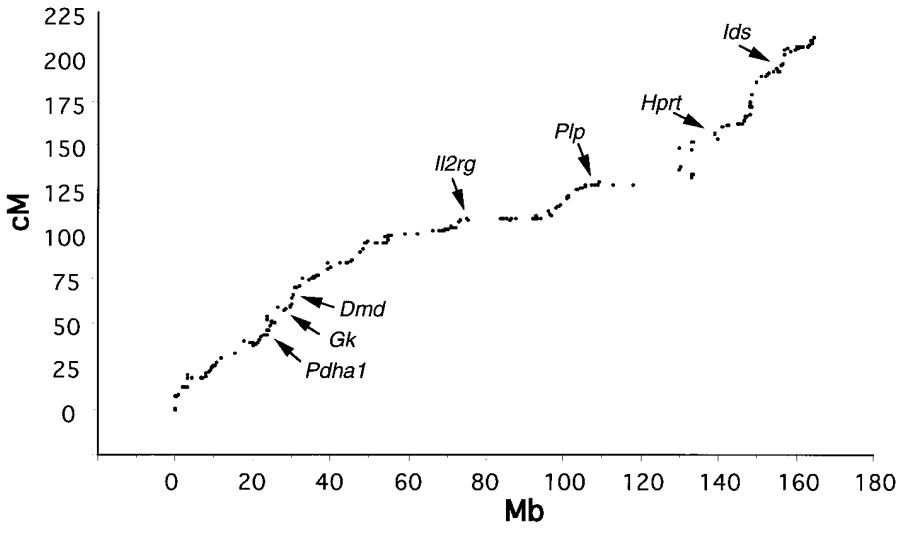

Figure 1.—Scatterplot of genetic (cM) *vs.* physical (Mb) map position for 249 loci on the human *X* chromosome (Wang *et al.* 1994). The locations of 7 loci surveyed for heterozygosity are indicated; genetic map positions for these loci are given in Table 2. Nagaraja *et al.* (1997) have also compared recombination rates in different regions of the *X* chromosome using genetic and physical maps integrated with 187 microsatellite markers. Their comparison, based on fewer markers, reveals regional variation in recombination rate consistent with that shown here. Physical and genetic distances from the *Xp* telomere are both expected to be nondecreasing functions. Irregularities, such as seen near *Hprt*, may be due to experimental errors in the physical map, variance in the genetic map, or both (Wang *et al.* 1994; Nagaraja *et al.* 1997).

TABLE 2

Polymorphisms in humans and divergence between human-chimpanzee and human-orangutan at introns of seven *X*-linked loci

Locus	Genetic map position (cM)	Recombination rate[a] (cM/Mb)	Length (bp)	Polymorphic sites	π (SE) (%)	θ (SE) (%)	Tajima's *D*	Fu and Li's *D*	Divergence *Homo-Pan* (%)	Divergence *Homo-Pongo* (%)
Il2rg	108.3	0.27	1,147	0	0.000 (0.000)	0.000 (0.000)	—	—	0.78	2.53
Plp	127.7	0.49	769	2	0.095 (0.087)	0.092 (0.076)	0.120	−0.410	0.65	1.30
Hprt	160.4	1.44	2,485	4	0.038 (0.032)	0.057 (0.038)	−1.245	−1.426	0.97	2.20
Gk	57.0	1.47	1,861	1	0.019 (0.023)	0.019 (0.021)	0.015	0.740	0.64	1.88
Ids	192.4	2.68	1,909	0	0.000 (0.000)	0.000 (0.000)	—	—	0.26	0.84
Pdha1	42.8	3.85	1,657	5	0.137 (0.093)	0.107 (0.067)	1.133	0.592	0.84	2.78
Dmd	63.6	7.74	1,537	8	0.187 (0.121)	0.184 (0.104)	0.059	−0.318	0.85	2.93
All loci	—	—	11,365	20	0.063 (0.036)	0.062 (0.031)	0.072	−0.349	0.72	2.08

[a] Loci arranged in rank order by recombination rate.

Figure 2.—Scatterplots of nucleotide diversity (π) *vs.* recombination rate (a), *Homo-Pan* divergence *vs.* recombination rate (b), and π / divergence *vs.* recombination rate (c). Regression line is indicated by a solid line.

Miller 1993), despite the fact that this approach includes multiple nonindependent comparisons. Instead, we followed the approach suggested by Lewontin (1995) and calculated the significance of D' using Fisher's exact test for a set of 11 independent pairwise comparisons among the 12 nonunique sites. Independent tests were constructed by taking the sites in order and comparing the pairs 1–2, 2–3, 3–4, . . . , 11–12. Three of the 11 comparisons revealed significant linkage disequilibrium. These 3 significant comparisons were between pairs of sites within *Pdha1*; no significant linkage disequilibrium was observed in comparisons between sites from different genes. The two major haplotypes at *Pdha1* are both present in a single geographic region (Namibia; Table 1), suggesting that the observed linkage disequilibrium is not simply a consequence of popu-

lation subdivision. Given the overall low level of diversity and small sample size, the power to detect linkage disequilibrium is quite low. It is noteworthy that the gene with the highest recombination rate, *Dmd*, was the only locus at which all four gametic types were observed between polymorphic sites. We calculated the neutral recombination parameter, γ, from the polymorphism data at *Dmd* using the method of Hey and Wakely (1997). This provided an estimate of the per nucleotide recombination rate, $c = 1.196 \times 10^{-7}$ (assuming $N_e = 10^4$; Hammer 1995), or 11.96 cM/Mb, compared to the value of 7.74 cM/Mb obtained experimentally from mapping data (Figure 1). We did not calculate γ for other loci because the small number of segregating sites is insufficient for a reasonable estimate.

The average level of nucleotide divergence between humans and chimpanzees was 0.72%, about one-half the value of 1.4% reported for fourfold degenerate sites in autosomes (Hammer 1995). This difference may be due to different levels of constraint on introns and fourfold degenerate sites, to male-driven molecular evolution, or to a lower mutation rate on the *X* relative to autosomes (McVean and Hurst 1997). Divergence ranged from a low value of 0.26% at *Ids* to a high value of 0.97% at *Hprt*. Nine insertion/deletion differences were observed between humans and chimpanzee. The average level of divergence between humans and orangutan was 2.08% and ranged from a low of 0.84% at *Ids* to a high of 2.98% at *Dmd*. Twenty-two insertion/deletion differences were observed between humans and orangutan. The observed divergence between humans and chimpanzee was used to calculate the neutral mutation rate (μ) for each gene under the assumption that observed differences are functionally equivalent. Assuming a divergence time of 5 million yr and a generation time of 20 yr, the average value was 1.44×10^{-8} (Table 3). This value is on the order of the estimate of the underlying mutation rate ($1–2 \times 10^{-8}$) for humans (Drake *et al.* 1998), suggesting that the introns surveyed here show relatively little functional constraint.

We performed a single, seven-locus HKA test of the neutral expectation of equal ratios of polymorphism to divergence among genes (Hudson *et al.* 1987). Despite the fact that polymorphism varied by more than one order of magnitude among the loci while divergence varied by less than a factor of four (Table 2), the HKA comparison did not reject the null model ($\chi^2 = 7.36$, d.f. = 5, $P > 0.10$).

A scatterplot of nucleotide diversity (π) *vs.* recombination rate is shown in Figure 2a. There is a significant positive correlation between nucleotide heterozygosity (either π or θ) and rate of recombination (for π, $R^2 = 0.600$, $P = 0.04$; for θ, $R^2 = 0.601$, $P = 0.04$). If this was due to differences in mutation rate or differences in the level of selective constraint among loci, we would expect to see a positive correlation between divergence (D) and recombination rate. However, there is no correlation between D and recombination rate, using *Homo-Pan* (Figure 2b) or *Homo-Pongo* divergences ($P > 0.25$ for both; data in Table 2). The significant correlation between heterozygosity and recombination rate is based on only seven data points and must be considered tentative. Indeed, if *Dmd* is removed from the analysis, the correlation is no longer significant ($R^2 = 0.189$, $P = 0.38$). One might expect that some of the scatter is due to differences in underlying mutation rate or level of constraint. If so, then a plot of π/D *vs.* recombination rate would be expected to show a tighter relationship than a plot of π *vs.* recombination rate. However, this is not the case (Figure 2c), suggesting that the unex-

TABLE 3

Mutation rate, effective population size, and coalescence time estimated from each of seven *X*-linked genes

Locus	μ[a]	N_e[b]	T_{mean} (yr)[c]	T_{mode} (yr)[c]	95% CI for T (yr)
Il2rg	1.56×10^{-8}	0	536,000	361,000	169,000–1,265,000
Plp	1.30×10^{-8}	24,400	864,000	583,000	274,000–1,997,000
Hprt	1.94×10^{-8}	6,500	717,000	536,000	266,000–1,511,000
Gk	1.28×10^{-8}	4,900	535,000	378,000	181,000–1,208,000
Ids	5.20×10^{-9}	0	449,000	314,000	148,000–1,028,000
Pdha1	1.68×10^{-8}	27,200	1,019,000	760,000	367,000–2,137,000
Dmd	1.70×10^{-8}	36,700	1,543,000	1,253,000	589,000–3,007,000
All loci	1.44×10^{-8}	14,600	743,000	655,000	380,000–1,258,000

Loci arranged in rank order by recombination rate, as in Table 2.

[a] Mutation rate (μ) per nucleotide per generation was calculated from the *Homo-Pan* divergence data in Table 2, assuming a divergence time of 5 mya and a generation time of 20 yr.

[b] Effective population size (N_e) was calculated from the neutral expectation for *X*-linked genes, $\pi = 3N_e\mu$, and sex ratio of one.

[c] Two estimates of coalescence time (T), T_{mean} and T_{mode}, calculated as in Fu and Li (1997), assuming $N_e = 14{,}600$ and $\mu = 1.44 \times 10^{-8}$. $T_{mean} = E(T|k_{max})$, where k_{max} is the maximum number of nucleotide differences between two sequences in a sample. T_{mode} maximizes the posterior probability $P(T|k_{max})$.

plained variance is not due simply to differences among loci in the neutral mutation rate.

DISCUSSION

Nucleotide diversity in humans: The average level of nucleotide diversity in humans for *X*-linked introns is clearly quite low (π = 0.063%). How does this value compare to previous estimates? To compare directly to variation at autosomal loci, π for *X*-linked loci must be multiplied by $^4/_3$ to account for differences in effective population size ($\pi_{standardized}$ = 0.063% × $^4/_3$ = 0.084%). Previous estimates of π for nuclear genes in humans are given in Table 4. Comparisons among studies are somewhat problematic because of different sampling strategies, including very large, worldwide samples at one extreme (Harding *et al.* 1997) and comparisons between two Caucasian alleles at the other extreme (Li and Sadler 1991). The level of constraint may also differ among the regions surveyed that include introns, noncoding DNA, and silent sites. Mutation rates may differ as well, either due to male-driven molecular evolution (Shimmin *et al.* 1993) or to *X*-autosome differences in mutation rate (McVean and Hurst 1997). Nonetheless, the average value reported here is in general agreement with previous studies.

The average value, however, masks the substantial variation observed among loci. The range of variation among the seven loci surveyed on the *X* chromosome (Table 2) is greater than the range of variation among all previous studies (Table 4). Nucleotide diversity at *Dmd* ($\pi_{standardized}$ = 0.187% × $^4/_3$ = 0.249%) is somewhat higher than observed at β-globin (π = 0.18%; Harding *et al.* 1997) or at *Lpl* (π = 0.20%; Clark *et al.* 1998; Nickerson *et al.* 1998), while two loci on the *X* (*Il2rg* and *Ids*) showed no variation, like the *Zfy* intron on the *Y* chromosome (Dorit *et al.* 1995). The average level of π in humans for silent and noncoding DNA (~0.1%) is about one order of magnitude less than the average in *D. melanogaster* for silent sites (~1.3%; Moriyama and Powell 1996). In both humans and *D. melanogaster*, the highest levels of nucleotide diversity at any locus are about 2.5 times the average for all loci (*Dmd*, 0.249% in humans from this study; *Amy-p*, 3.3% in *D. melanogaster*; Moriyama and Powell 1996). Thus, the range of variation in heterozygosity among loci is comparable in both species.

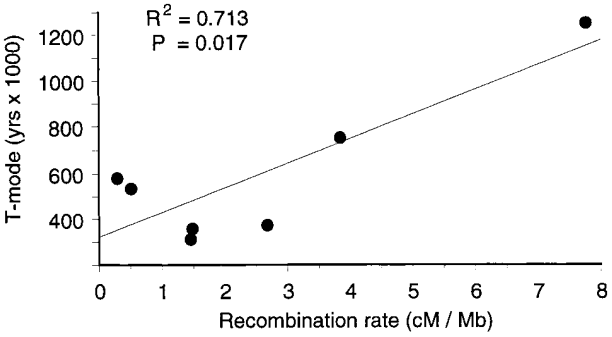

Figure 3.—Scatterplot of coalescence time (T_{mode}) *vs.* recombination rate.

We have used the mutation rates estimated from divergence data and the level of heterozygosity estimated from polymorphism data to calculate the effective population size (N_e) and coalescence time (T_{mean} and T_{mode}) for each locus (Table 3). The variation in heterozygosity among loci in conjunction with the rough constancy of mutation rate implies that effective population size and coalescence time vary among loci. Indeed, there is a positive correlation between the regional rate of recombination and the coalescence time for the seven loci (Figure 3). The locus with lowest heterozygosity at which polymorphisms were observed (*Gk*) gives an estimate of N_e = 4900 individuals and T_{mode} = 378,000 yr, while estimates for *Dmd* are N_e = 36,700 individuals and

TABLE 4

Nucleotide diversity from nuclear genes in humans

Locus[a]	Chromosome	Sample size	Length (bp)	π (%)	$\pi_{standardized}$ (%)[b]	Reference
49 loci	Autosomes	2	8,537	0.110	0.110	Li and Sadler (1991)
Zfy intron	*Y*	38	729	0.000	0.000	Dorit *et al.* (1995)
YAP region	*Y*	16	2,638	0.037	0.148	Hammer (1995)
Sry region	*Y*	5	18,300	0.008	0.031	Whitfield *et al.* (1995)
β-globin	11	349	2,670	0.180	0.180	Harding *et al.* (1997)
Lpl	8	142	9,700	0.200	0.200	Nickerson *et al.* (1998)
Pdha1 intron	*X*	8	1,769	0.113	0.151	Hey (1997)
Zfx intron	*X*	29	1,151	0.040	0.053	Huang *et al.* (1998)
7-loci introns	*X*	10	11,365	0.063	0.084	This study

[a] The data from Li and Sadler (1991) are for fourfold degenerate sites, the data of Hammer (1995) are for a noncoding region of the *Y*, the data of Whitfield *et al.* (1995) are from a region of noncoding DNA flanking the *Sry* locus, and the data of Harding *et al.* (1997) and Nickerson *et al.* (1998) include mostly introns and flanking sequence but also include exons.

[b] Nucleotide diversity values for the *Y* chromosome were multiplied by 4 and values for the *X* were multiplied by $^4/_3$ to be directly comparable to values from autosomes.

T_{mode} = 1,253,000 yr. While the number of individuals in the population is the same regardless of the locus sampled, the number of individuals effectively contributing to heterozygosity may differ among loci as a consequence of selection on linked sites (see below). Consequently, rates of adaptive and deleterious evolution may vary among genes with different effective population sizes. Rates of adaptive evolution are expected to be greatest in regions with large effective population size, while fixation rates for deleterious mutations are expected to be highest in genomic regions with small effective population size.

Tests of selection: Both Tajima's *D* and Fu and Li's *D* are expected to be positive when there is an excess of intermediate-frequency sites. Hey (1997) compared Tajima's *D* values for mitochondrial loci and three nuclear loci (β-globin, elastin, and *Pdha1*) in humans. While none of the nuclear genes showed values of Tajima's *D* that were significantly greater than zero, all were positive and were significantly different from the negative values obtained from mtDNA. Hey argued that this difference between nuclear and mitochondrial genes is not consistent with a simple demographic model and is best explained by different evolutionary forces acting on nuclear and mitochondrial loci. The data presented here do not support the idea that nuclear genes, in general, show a tendency toward intermediate-frequency variants. None of the values of Tajima's *D* or Fu and Li's *D* were significant, and the values for both statistics with all the data considered together are close to the neutral expectation of zero. For Tajima's *D*, one locus was slightly negative (*Hprt*), one slightly positive (*Pdha1*), and three were very close to zero (*Plp*, *Gk*, and *Dmd*). For Fu and Li's *D*, which is sensitive to singletons, three loci were slightly negative (*Plp*, *Hprt*, and *Dmd*) and two were slightly positive (*Gk*, *Pdha1*). The power of these tests is quite low with small samples (Braverman *et al.* 1995; Simonsen *et al.* 1995), but there is no evidence for selection in these analyses. The one way in which these results differ from mitochondrial data is that nearly all samples of mitochondrial loci in humans reveal an excess of low-frequency variants (*e.g.*, Whittam *et al.* 1986; Excoffier 1990; Nachman *et al.* 1996; Hey 1997).

A single, seven-locus HKA test failed to reject the null model of equal proportions of polymorphism to divergence among loci. While there may be hints of nonneutral patterns in the observed ratios of polymorphism to divergence, larger datasets will be required before statistical tests of neutrality based on polymorphism data will provide much power to detect the signature of selection using this approach. Future efforts should be directed toward sequencing more bases rather than more individuals because this will increase the number of polymorphic sites most efficiently. It is clear that, if they exist, documenting the sort of nonneutral patterns that have been found in *D. melanogaster* (*e.g.*, Hudson *et al.* 1987; McDonald and Kreitman 1991) will require a much greater effort in humans.

Heterozygosity and rates of recombination: The scatterplots presented in Figure 2 provide another way to look for the footprint of selection and suggest that there is a positive correlation between heterozygosity and recombination rate yet no correlation between divergence and recombination rate. This conclusion must be treated as tentative because it is based on only seven data points. Moreover, the plot of π/D *vs.* recombination rate (Figure 2c) is not significant ($P = 0.08$). Nonetheless, there is a clear trend, and that trend is consistent with observations in other species (Begun and Aquadro 1992, 1993; Aguade and Langley 1994; Aquadro *et al.* 1994; Stephan 1994; Moriyama and Powell 1996; Nachman 1997).

Two models have been proposed to explain a correlation between heterozygosity and recombination rate. Background selection refers to the removal of deleterious mutations from a population and the associated removal of linked neutral variation (Charlesworth *et al.* 1993, 1995; Charlesworth 1994; Hudson 1994; Hudson and Kaplan 1994, 1995). The strength of this effect depends on the deleterious mutation rate, the average selection coefficient and dominance factor, and the frequency of recombination. Background selection is a neutral model as it invokes only neutral and deleterious mutations. Genetic hitchhiking refers to the fixation of adaptive mutations and the associated fixation of linked neutral variation (Maynard Smith and Haigh 1974; Kaplan *et al.* 1989; Wiehe and Stephan 1993; Stephan 1995). During the course of a selective sweep, variability is eliminated from the population in a genomic region around the nucleotide under selection, and the size of this region depends in part on the rate of recombination. Both genetic hitchhiking and background selection are based on the joint effects of selection and linkage, although they have different biological implications. If genetic hitchhiking is the predominant force producing the observed pattern, this implies that adaptive evolution at the molecular level is common. Background selection, in contrast, invokes no special role for adaptive mutations.

One potential means for discriminating between these contrasting models lies in the frequency distribution of polymorphic sites: strong selective sweeps are expected to create a skew in the frequency distribution while background selection is typically expected to leave a neutral, equilibrium distribution (Charlesworth 1994). The observed values of Tajima's *D* and Fu and Li's *D* test statistics provide no evidence for a skew in the frequency distribution of polymorphisms (Table 2) and as such are consistent with background selection. However, as stated above, the power of these tests is low with small samples and thus the present data do not provide a strong means of discriminating between adap-

tive and deleterious mutations as the cause of the correlation.

Are there biological differences between *D. melanogaster* and humans that might, *a priori*, lead one to expect that selection might be more effective in one species than the other in producing a correlation between heterozygosity and recombination? The total genetic map in *D. melanogaster* is 277 cM, and because there is no recombination in males, the effective amount of recombination is 138.5 cM. There are ~12,000–16,000 genes in Drosophila (Bird 1995). Considering only genes, we can express the density of targets for selection in terms of recombinational distance as 14,000 genes/138.5 cM = 101 genes/cM. The sex-averaged human genetic map is 3699 cM (Dib *et al.* 1996) and humans have ~70,000 genes (Bird 1995), giving a density of 19 genes/cM. Thus, the targets for selection in humans may be approximately one-fifth as dense as in Drosophila. Of course, there may be targets for selection other than the nucleotides that constitute genes, but it seems reasonable that the number of targets for both deleterious mutations and for advantageous mutations may be fewer for the same recombinational distance in humans than in *D. melanogaster*.

Human evolution: The variation among loci in levels of diversity has implications for understanding human evolution. The correlation between heterozygosity and recombination rate implies that genes in regions of highest recombination are least likely to be perturbed from a neutral equilibrium state by the effects of selection at linked sites. Thus, these genes should provide the best estimates of neutral parameters such as N_e and coalescence time. There are two hotspots for recombination in the *Dmd* gene, one in intron 7 and one in intron 44, flanking the region we have surveyed (Oudet *et al.* 1992). Regions surrounded by hotspots of recombination may be largely uncoupled from the effects of selection elsewhere on the chromosome. The two loci from this study with the highest recombination rates (*Pdha1* and *Dmd*) lead to estimates of N_e for humans on the order of 30,000 individuals and coalescence times of more than a million years. Similarly, Harding *et al.* (1997) estimated a coalescence time of 1,100,000 yr for the β-globin locus. This falls near the upper end of the range we have documented here (once *X*-autosome corrections are made), comparable to loci with medium to high rates of recombination. There is a hotspot for recombination in the 5′ region of the β-globin locus, but the recombinational environment to the other side of the gene is unknown.

The positive correlation between coalescence times and recombination rate for the seven loci we have surveyed (Figure 3) may help reconcile much of the heterogeneity in currently available estimates of N_e and coalescence time from levels of sequence variation of different gene regions. Genes in different genomic regions clearly have different evolutionary histories and will thus provide different pictures of human evolutionary history, with genes in regions of highest recombination providing the deepest views into our genetic heritage.

We thank Y. X. Fu for help in calculating coalescence times, M. F. Hammer for discussions and human DNAs, and O. A. Ryder for chimpanzee and orangutan DNAs. This work was supported by the National Science Foundation and the National Institutes of Health.

LITERATURE CITED

Aguade, M., and C. H. Langley, 1994 Polymorphism and divergence in regions of low recombination in Drosophila, pp. 67–76 in *Non-Neutral Evolution: Theories and Molecular Data*, edited by B. Golding. Chapman & Hall, New York.

Aquadro, C. F., 1992 Why is the genome variable? Insights from Drosophila. Trends Genet. **8:** 355–362.

Aquadro, C. F., 1997 Insights into the evolutionary process from patterns of DNA sequence variability. Curr. Opin. Genet. Devel. **7:** 835–840.

Aquadro, C. F., D. J. Begun and E. C. Kindahl, 1994 Selection, recombination, and DNA polymorphism in Drosophila, pp. 46–56 in *Non-Neutral Evolution: Theories and Molecular Data*, edited by B. Golding. Chapman & Hall, New York.

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

Begun, D. J., and C. F. Aquadro, 1993 African and North American populations of Drosophila melanogaster are very different at the DNA level. Nature **365:** 548–550.

Bird, A. P., 1995 Gene number, noise reduction and biological complexity. Trends Genet. **11:** 94–100.

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 polymorphism. Genetics **140:** 783–796.

Charlesworth, B., 1994 The effect of background selection against deleterious mutations on weakly selected, linked variants. Genet. Res. **63:** 213–227.

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.

Clark, A. G., K. M. Weiss, D. A. Nickerson, S. L. Taylor, A. Buchanan *et al.*, 1998 Haplotype structure and population genetic inferences from nucleotide sequence variation in human lipoprotein lipase. Am. J. Hum. Genet. **63:** 595–612.

Dib, C., S. Faure, C. Fizames, D. Samson, N. Drouot *et al.*, 1996 A comprehensive genetic map of the human genome based on 5,264 microsatellites. Nature **380:** 152–154.

Diehl, H. J., M. Schaich, R. M. Budzinski and W. Stoffel, 1986 Individual exons encode the integral membrane domains of human myelin proteolipid protein. Proc. Natl. Acad. Sci. USA **83:** 9807–9811.

Dorit, R. L., H. Akashi and W. Gilbert, 1995 Absence of polymorphism at the *Zfy* locus on the human Y chromosome. Science **268:** 1183–1185.

Drake, J. W., B. Charlesworth, D. Charlesworth and J. F. Crow, 1998 Rates of spontaneous mutation. Genetics **148:** 1667–1686.

Edwards, A., H. Voss, P. Rice, A. Civitello, J. Stegemann *et al.*, 1990 Automated DNA sequencing of the human HPRT locus. Genomics **6:** 593–608.

Excoffier, L., 1990 Evolution of human mitochondrial DNA: evidence for departure from a pure neutral model of populations at equilibrium. J. Mol. Evol. **30:** 125–139.

Fu, Y. X., 1996 Estimating the age of the common ancestor of a DNA sample using the number of segregating sites. Genetics **144:** 829–838.

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

Fu, Y. X., and W. H. Li, 1997 Estimating the age of the common

ancestor of a sample of DNA sequences. Mol. Biol. Evol. **14:** 195–199.

Fullerton, S. M., 1996 Allelic sequence variation at the human β-globin locus, pp. 225–241 in *Molecular Biology and Human Diversity*, edited by A. J. Boyce and C. G. N. Mascie-Taylor. Cambridge University Press, Cambridge, UK.

Fullerton, S. M., R. M. Harding, A. J. Boyce and J. B. Clegg, 1994 Molecular and population genetic analysis of allelic sequence diversity at the human β-globin locus. Proc. Natl. Acad. Sci. USA **91:** 1805–1809.

Guo, W., K. Worley, V. Adams, J. Mason, D. Sylvester-Jackson *et al.*, 1993 Genomic scanning for expressed sequences in Xp21 identifies the glycerol kinase gene. Nat. Genet. **4:** 367–372.

Hammer, M., 1995 A recent common ancestry for human Y chromosomes. Nature **378:** 376–378.

Harding, R. M., S. M. Fullerton, R. C. Griffiths, J. Bond, M. J. Cox *et al.*, 1997 Archaic African *and* Asian lineages in the genetic ancestry of modern humans. Am. J. Hum. Genet. **60:** 772–789.

Hartl, D. L., and A. G. Clark 1997 *Principles of Population Genetics, Ed. 3*, Sinauer Associates, Inc., Sunderland, MA.

Hey, J., 1997 Mitochondrial and nuclear genes present conflicting portraits of human origins. Mol. Biol. Evol. **14:** 166–172.

Hey, J., and J. Wakeley, 1997 A coalescent estimator of the population recombination rate. Genetics **145:** 833–846.

Huang, W., Y. X. Fu, B. H. J. Chang, X. Gu, L. B. Jorde *et al.*, 1998 Sequence variation in Zfx introns in human populations. Mol. Biol. Evol. **15:** 138–142.

Hudson, R. R., 1994 How can the low levels of DNA sequence variation in regions of the Drosophila genome with low recombination rates be explained? Proc. Natl. Acad. Sci. USA **91:** 6815–6818.

Hudson, R. R., and N. L. Kaplan, 1994 Gene trees with background selection, pp. 140–153 in *Non-Neutral Evolution: Theories and Molecular Data*, edited by B. Golding. Chapman and Hall, New York.

Hudson, R. R., and N. L. Kaplan, 1995 Deleterious background selection with recombination. Genetics **141:** 1605–1617.

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

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

Koike, K., Y. Urata, S. Matsuo and M. Koike, 1990 Characterization and nucleotide sequence of the gene encoding the human pyruvate dehydrogenase α-subunit. Gene **93:** 307–311.

Kreitman, M., 1983 Nucleotide polymorphism at the alcohol dehydrogenase locus of *Drosophila melanogaster*. Nature **304:** 412–417.

Kreitman, M., and H. Akashi, 1995 Molecular evidence for natural selection. Annu. Rev. Ecol. Syst. **26:** 403–422.

Lewontin, R. C., 1964 The interaction of selection and linkage. I. General considerations: heterotic models. Genetics **43:** 419–434.

Lewontin, R. C., 1995 The detection of linkage disequilibrium in molecular sequence data. Genetics **140:** 377–388.

Li, W. H., and L. A. Sadler, 1991 Low nucleotide diversity in man. Genetics **129:** 513–523.

Lu, F., J. Lu, R. L. Clingan, M. A. Wentland, D. M. Munzy *et al.*, 1994 Complete DNA sequence of the human iduronate sulphate sulphatase (*Ids*) locus. Unpublished. GenBank accession no. L35485.

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

McDonald, J. H., and M. Kreitman, 1991 Adaptive protein evolution at the Adh locus in *Drosophila*. Nature **351:** 652–654.

McVean, G. T., and L. D. Hurst, 1997 Evidence for a selectively favourable reduction in the mutation rate of the X chromosome. Nature **386:** 388–392.

Moriyama, E. N., and J. R. Powell, 1996 Intraspecific nuclear DNA variation in Drosophila. Mol. Biol. Evol. **13:** 261–277.

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

Nachman, M. W., W. M. Brown, M. Stoneking and C. F. Aquadro, 1996 Nonneutral mitochondrial DNA variation in humans and chimpanzees. Genetics **142:** 953–963.

Nagaraja, R., S. MacMillan, J. Kere, C. Jones, S. Griffin *et al.*, 1997 X chromosome map at 75-kb STS resolution, revealing extremes of recombination and GC content. Genome Res. **7:** 210–222.

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

Nickerson, D. A., S. L. Taylor, K. M. Weiss, A. G. Clark, R. G. Hutchinson *et al.*, 1998 DNA sequence diversity in a 9.7 kb region of the human lipoprotein lipase gene. Nat. Genet. **19:** 233–240.

Noguchi, M., S. Adelstein, X. Cao and W. J. Leonard, 1993 Characterization of the human interleukin-2 receptor γ chain gene. J. Biol. Chem. **268:** 13601–13608.

Oudet, C., A. Hanauer, P. Clemens, T. Caskey and J. L. Mandel, 1992 Two hot spots of recombination in the DMD gene correlate with the deletion prone regions. Hum. Mol. Genet. **1:** 599–603.

Richards, S., 1992 Sequence analysis of a deletion hotspot in intron 44 of the Dystrophin gene. Unpublished. GenBank accession no. M86524.

Roest Crollius, H., M. T. Ross, A. Grigoriev, C. J. Knights, E. Holloway *et al.*, 1996 An integrated YAC map of the human X chromosome. Genome Res. **6:** 943–955.

Saiki, R. K., D. H. Gelfand, S. Stoffel, S. J. Scharf, R. Higuchi *et al.*, 1988 Primer-directed enzymatic amplification of DNA with a thermostable DNA polymerase. Science **239:** 487–491.

Sanger, F., S. Nicklon and A. R. Coulson, 1977 DNA sequencing with chain terminating inhibitors. Proc. Natl. Acad. Sci. USA **74:** 5463–5467.

Schaeffer, S. W., and E. L. Miller, 1993 Estimates of linkage disequilibrium and the recombination parameter determined from segregating nucleotide sites in the alcohol dehydrogenase region of Drosophila pseudoobscura. Genetics **135:** 541–552.

Shimmin, L. C., B. H. J. Chang and W. H. Li, 1993 Male-driven evolution of DNA sequences. Nature **362:** 745–747.

Simonsen, K. L., G. A. Churchill and C. F. Aquadro, 1995 Properties of statistical tests of neutrality for DNA polymorphism data. Genetics **141:** 413–439.

Stephan, W., 1994 Effects of genetic recombination and population subdivision on nucleotide sequence variation in *Drosophila ananassae*, pp. 57–66 in *Non-Neutral Evolution: Theories and Molecular Data*, edited by B. Golding. Chapman and Hall, New York.

Stephan, W., 1995 An improved method for estimating the rate of fixation of favorable mutations based on DNA polymorphism data. Mol. Biol. Evol. **12:** 959–962.

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

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

Wang, L. H., A. Collins, S. Lawrence, B. J. Keats and N. E. Morton, 1994 Integration of gene maps: chromosome X. Genomics **22:** 590–604.

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

Whitfield, L. S., J. E. Sulston and P. E. Goodfellow, 1995 Sequence variation of the human Y chromosome. Nature **378:** 379–380.

Whittam, T. S., A. G. Clark, M. Stoneking, R. L. Cann and A. C. Wilson, 1986 Allelic variation in human mitochondrial genes based on patterns of restriction site polymorphism. Proc. Natl. Acad. Sci. USA **83:** 9611–9615.

Wiehe, T. H. E., and W. Stephan, 1993 Analysis of a genetic hitchhiking model, and its application to DNA polymorphism data from *Drosophila melanogaster*. Mol. Biol. Evol. **10:** 842–855.

Communicating editor: A. G. Clark