Contrasting Evolutionary Histories of Two Introns of the Duchenne Muscular Dystrophy Gene, *Dmd*, in Humans

Michael W. Nachman and Susan L. Crowell

Department of Ecology and Evolutionary Biology, University of Arizona, Tucson, Arizona 85721

Manuscript received September 13, 1999
Accepted for publication April 17, 2000

ABSTRACT

The Duchenne muscular dystrophy (*Dmd*) locus lies in a region of the X chromosome that experiences a high rate of recombination and is thus expected to be relatively unaffected by the effects of selection on nearby genes. To provide a picture of nucleotide variability at a high-recombination locus in humans, we sequenced 5.4 kb from two introns of *Dmd* in a worldwide sample of 41 alleles from Africa, Asia, Europe, and the Americas. These same regions were also sequenced in one common chimpanzee and one orangutan. Dramatically different patterns of genetic variation were observed at these two introns, which are separated by >500 kb of DNA. Nucleotide diversity at intron 44 ($\pi = 0.141\%$) was more than four times higher than nucleotide diversity at intron 7 ($\pi = 0.034\%$) despite similar levels of divergence for these two regions. Intron 7 exhibited significant linkage disequilibrium extending over 10 kb and also showed a significant excess of rare polymorphisms. In contrast, intron 44 exhibited little linkage disequilibrium and no skew in the frequency distribution of segregating sites. Intron 7 was much more variable in Africa than in other continents, while intron 44 displayed similar levels of variability in different geographic regions. Comparison of intraspecific polymorphism to interspecific divergence using the HKA test revealed a significant reduction in variability at intron 7 relative to intron 44, and this effect was most pronounced in the non-African samples. These results are best explained by positive directional selection acting at or near intron 7 and demonstrate that even genes in regions of high recombination may be influenced by selection at linked sites.

IDENTIFYING the forces shaping genetic variation in natural populations remains a key problem in population genetics. Surprisingly, our understanding of the amount and structure of genetic variation at the nucleotide level in humans is still in its early stages. Mutation, migration, drift, recombination, selection at individual loci, the effects of selection at linked sites, and demographic history undoubtedly all play a role in shaping patterns of human genetic variation, although the relative importance of these different factors is not yet clear. Significant progress into this problem has been made with recent studies of nucleotide variation at β-globin (Harding *et al.* 1997), dystrophin (Zietkiewicz *et al.* 1997, 1998), lipoprotein lipase (Clark *et al.* 1998), introns of seven X-linked genes (Nachman *et al.* 1998), pyruvate dehydrogenase E1 α subunit (Harris and Hey 1999), angiotensin converting enzyme (Rieder *et al.* 1999), a noncoding region at Xq13.3 (Kaessmann *et al.* 1999), the X-chromosome-specific zinc-finger protein (Jaruzelska *et al.* 1999), and the melanocortin 1 receptor (Rana *et al.* 1999; Harding *et al.* 2000). Collectively, these studies have shown that the average level of nucleotide diversity in humans is quite low, largely confirming a result first obtained by Li and Sadler (1991). However, there is also substantial heterogeneity in levels and patterns of genetic variation among loci, and a central challenge now is to explain these differences.

Theoretical studies show that the interaction of selection and recombination can have a dramatic effect on levels of nucleotide variability, either through the fixation of advantageous mutations (*i.e.*, genetic hitchhiking; Maynard Smith and Haigh 1974) or the removal of deleterious mutations (*i.e.*, background selection; Charlesworth *et al.* 1993). Both processes are expected to reduce levels of neutral genetic variation in genomic regions with low rates of recombination. Estimates of recombination rate for different genomic regions can be obtained by comparing the genetic and physical locations of markers. In humans, there is evidence for both local and large-scale variation in the recombinational landscape. For example, several studies have revealed recombinational hotspots, suggesting that recombination rates may vary substantially over a scale of several kilobases (*e.g.*, Oudet *et al.* 1992; Harding *et al.* 1997). A common large-scale pattern is the suppression of recombination near centromeres of metacentric chromosomes (*e.g.*, Nagaraja *et al.* 1997). Variation at both of these scales is likely to be important in determining the effects of selection at linked sites. There is good evidence for a positive correlation be-

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

tween regional recombination rate and levels of nucleotide heterozygosity in *Drosophila melanogaster* (Begun and Aquadro 1992; Aquadro *et al.* 1994; Moriyama and Powell 1996) and weaker evidence for a positive association between recombination rate and levels of nucleotide variability in a number of other organisms (Nachman 1997; Dvorák *et al.* 1998; Kraft *et al.* 1998; Stephan and Langley 1998), including humans (Nachman *et al.* 1998; Przeworski *et al.* 2000).

Motivated by theoretical expectations concerning the effects of selection on linked neutral variation and the empirical evidence suggesting that such effects may be common, we were interested in documenting patterns of nucleotide variability at a gene that experiences a very high rate of recombination in humans. In principle, high-recombination genes are least likely to be affected by selection at linked sites and are thus more likely to reflect neutral, equilibrium conditions.

Dystrophin is the protein product of the Duchenne muscular dystrophy (*Dmd*) locus. Duchenne muscular dystrophy is a common inherited disease with an incidence worldwide of 1 in 3500 births, many of which arise from new mutations. The *Dmd* locus is ~2.4 Mb long and consists of 79 exons that encode a 14-kb transcript. This mRNA codes for a 3685-amino-acid protein of 427 kD that shows similarity to several cytoskeletal proteins. *Dmd* is X-linked and lies in a genomic region experiencing high rates of recombination. Fine scale mapping of this region reveals overall recombination frequencies of 12 cM across 2 Mb of DNA (Abbs *et al.* 1990; Oudet *et al.* 1992). This overall rate, 6 cM/Mb, is about six times the average value of ~1 cM/Mb across the human genome. Oudet *et al.* (1992) documented considerable heterogeneity in recombination frequencies in different regions of the *Dmd* gene and found that some regions experience recombination rates >10 cM/Mb (Figure 1). Previous studies have surveyed worldwide variation in intron 44 of *Dmd* using single-strand conformation polymorphisms (SSCPs) to detect mutations (Zietkiewicz *et al.* 1997, 1998) or through direct DNA sequencing (Nachman *et al.* 1998). Zietkiewicz *et al.* (1997, 1998) screened 7622 bp in a worldwide sample of 250 chromosomes but may not have uncovered all of the underlying variation since their study was based on mutation detection using SSCP. Nachman *et al.* (1998) surveyed 1537 bp in 10 individuals by sequencing all sites.

Here, we further investigate patterns of genetic variation at two introns (7 and 44) of *Dmd* in a global sample of 41 alleles and find strikingly different patterns of genetic variation in each region. Both of these introns experience recombination rates well above the genomic average and are expected to be relatively free of the effects of selection at linked sites. Nonetheless, the con-

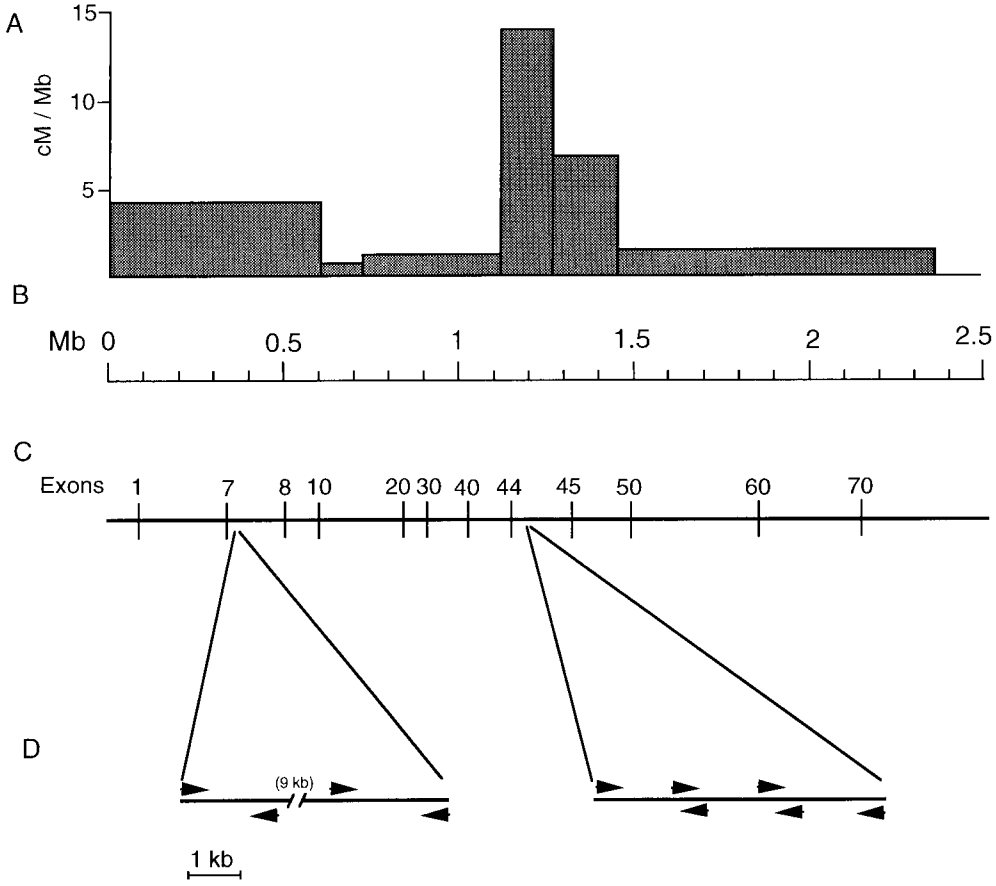

Figure 1.—Map of *Dmd.* (A) Recombination rate estimates for different portions of the *Dmd* locus (data from Oudet *et al.* 1992). (B) Physical map of *Dmd.* (C) Position of exons. (D) Amplified regions of introns 7 and 44 in this study. Arrows denote PCR primer positions.

trasting patterns of variation at these two introns suggest that recent directional selection has acted at or near intron 7 of *Dmd*.

MATERIALS AND METHODS

Samples: Forty-one men were sampled, including 10 from Africa, 10 from Europe, 11 from Asia (including one from Melanesia), and 10 from the Americas. Human genomic DNAs were provided by Dr. M. F. Hammer from the Y chromosome consortium (YCC) DNA repository. A single male common chimpanzee (*Pan troglodytes*) and a single male orangutan (*Pongo pygmaeus*) were also surveyed from DNAs provided by Dr. O. A. Ryder. By sequencing X chromosomes in males, we were able to amplify by PCR and sequence a single allele per individual and thus avoid problems associated with sequencing and scoring heterozygous sites. We were also able to recover haplotypes directly and thereby look at patterns of linkage disequilibrium among all sites in the sample.

PCR amplification and sequencing of *Dmd*: A map of the *Dmd* locus is shown in Figure 1. Additional detailed information about the structure of this locus can be found at http://www.dmd.nl. Intron 7 and intron 44 are separated by >500 kb of DNA. Both introns lie in genomic regions experiencing high rates of recombination (>4 cM/Mb), although the intervening introns experience considerably lower rates of recombination (<1 cM/Mb). DNA was PCR amplified (Saiki *et al.* 1988) in 25-μl volumes with 40 cycles of 94° 1 min, 55° 1 min, and 72° 2 min. Amplification primers were designed from published sequence for intron 7 (GenBank accession no. U60822) and intron 44 (GenBank accession no. M86524) and are listed in Table 1. Products were cycle-sequenced and run on an ABI 377 automated sequencer or sequenced manually as in Nachman *et al.* (1998). A total of 2389 bp from intron 7 and 3000 bp from intron 44 were sequenced. The 3000 bp in intron 44 are contiguous but the 2389 bp in intron 7 consist of two fragments (1388 bp and 1001 bp) separated by 8820 bp of largely repetitive DNA, which we found difficult to sequence (Figure 1). The 3000-bp portion of intron 44 includes and extends the smaller region (1537 bp) sequenced in 10 individuals in Nachman *et al.* (1998). Sequences have been submitted to GenBank under accession nos. AF279921–AF280049.

Data analysis: Sequences were aligned by eye, and the numbers and frequencies of all polymorphisms were counted. Two measures of nucleotide variability, π (Nei and Li 1979) and θ (Waterson 1975), were calculated. 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. Under neutral, equilibrium conditions, both π and θ estimate the parameter $3N_e\mu$ for X-linked loci, where N_e is the effective population size and μ is the neutral mutation rate. Departures from a neutral equilibrium frequency distribution of polymorphisms were evaluated using two approaches (Tajima 1989; Fu and Li 1993). Linkage disequilibrium (D') was calculated for a set of independent pairwise comparisons between nonunique polymorphic sites (Lewontin 1964, 1995), and the significance of D' was assessed using Fisher's exact tests. Ratios of polymorphism within humans to divergence between human and chimpanzee or human and orangutan were compared with expectations under a neutral model using the Hudson, Kreitman and Aguadé (HKA) test (Hudson *et al.* 1987). Polymorphism was based on variation segregating among the 41 human alleles and divergence was based on a single randomly chosen human allele and the single chimpanzee or orangutan allele.

RESULTS

Polymorphic sites for introns 7 and 44 are shown in Tables 2 and 3, respectively. Numbers of segregating sites, nucleotide diversity, measures of the frequency distribution, and levels of divergence are summarized in Table 4 for both introns. Nine segregating sites were observed in intron 7, and 19 segregating sites were observed in intron 44. Intron 7 had three insertion-deletion polymorphisms; two consisted of a single nucleotide and one consisted of 5 bp. Intron 44 contained a complicated compound microsatellite consisting of several different dinucleotide repeats (Table 3). Nucleotide diversity at intron 44 (π = 0.141%) was more than four times greater than nucleotide diversity at intron 7 (π = 0.034%). Waterson's θ, which is based on the number of segregating sites, was less than twice as large in intron 44 (θ = 0.148) as in intron 7 (θ = 0.088). The relative similarity in θ despite the difference in π

TABLE 1

Amplification primers used in this study

Intron	Direction	Primer
7 (section 1)	Forward	5′-TGT ATG CCC TAT GGA TGG AG-3′
	Reverse	5′-TCT GAT TAT GAT GTG GAT GC-3′
7 (section 2)	Forward	5′-TTT CCG CCT CTG CTG TAA TC-3′
	Reverse	5′-TTG TGA AAA TAA TCC TGG TAA G-3′
44	Forward (F1)	5′-TTG GGG GAA ATA AGG CAA TTC C-3′
	Reverse (R1)	5′-AGA AGC AAA AGA AGA ACC CCG C-3′
	Forward (F4)	5′-CTG TAG TAC CAG AAT CTC C-3′
	Reverse (R4)	5′-CTT GAA GTT GGG GAC ACT AGA G-3′
	Forward (F11)	5′-ATA ATG AGA ACT ACA AAC CAA G-3′
	Reverse (R5)	5′-AAC TCT GAA ATC CTA AAC AAA T-3′

Approximate positions of primers are shown in Figure 1. Amplification of 3 kb of intron 44 was done with three overlapping fragments using primer pairs F1 + R1, F4 + R4, and F11 + R5.

TABLE 2

Individuals sampled and polymorphic sites at *Dmd* intron 7

Continent	Country	Ethnic/language group	YCC no.	0000001 00000 1225562 45679 4261589 84312 5580189 10415
			Consensus	ATAA-TT A-TGG
Africa	Namibia	Tsumkwe	38	..... TA...
	Namibia	Tsumkwe	22	..G.... ....A
	South Africa	Sotho	32	....
	South Africa	Pedi	33	....A.
	South Africa	Herero	40	....
	C. African Repub.	Biaka	7	.GG.G.. TA...
	C. African Repub.	Biaka	6	..G.G.. TAC..
	Zaire	Mbuti	8	..G.G.. TA...
	Zaire	Mbuti	65	....
	Zaire	Mbuti	9	..G.G[a] . TA...
Europe	United Kingdom	United Kingdom	26	....
	Germany	German	61	....
	Germany	German	64	....
	Germany	German	62	....
	E. Europe	Ashkenazi	24	....
	Poland/Ukraine	Ashkenazi	59	....
	S. W. Russia	Adygean	56	.....C .....
	Russia	Russian	72	....
	Russia	Russian	71	....
	Turkey	Turkish	79	....
Asia	Japan	Japanese	78	....
	Japan	Japanese	76	....
	Japan	Japanese	77	....
	China	S. Han	66	....
	China	S. Han	67	....
	China	S. Han	68	....
	Cambodia	Cambodian	69	....
	Pakistan	Pakistani	57	....
	Siberia	Yakut	49	....
	Siberia	Yakut	51	....
	Melanesia	Nasioi	10	....
Americas	United States	Porch Creek	27	..G.... .....
	United States	Tohono O'Odham	1	G..... .....
	United States	Navajo	23	....
	United States	Amerindian	2	....
	United States	Amerindian	4	....
	Mexico	Mayan	17	....
	Brazil	Karitiana	12	...G... .....
	Brazil	Karitiana	13	....
	Brazil	Surui	16	....
	Brazil	Surui	14	....
			Pan	..G.G.. C....
			Pongo	..G.G.C T..A.

[a] Deletion of TTAAG.

between the two introns is due in large part to the difference in the number of singletons in each intron. Seven out of 9 (78%) polymorphic sites in intron 7 are singletons, while 6 out of 19 (32%) polymorphic sites in intron 44 are singletons. The frequency distribution of polymorphisms is consistent with neutral expectations for intron 44, but there is an excess of rare polymorphisms in intron 7, reflected in the significantly negative values of Tajima's *D* and Fu and Li's *D* statistics (Table 4).

Divergence was significantly higher at intron 7 than at intron 44 in comparisons between human and chimpanzee ($t = 2.30$, $P < 0.05$). In comparisons between human and orangutan, divergence was only slightly and not significantly higher at intron 7 than at intron 44 (Table 4).

TABLE 3

Polymorphic sites at *Dmd* intron 44

Continent	Country	YCC no.	0000000111 1 1 1 1 1 1 1 1222222 0167889123 7 7 8 8 8 8 8 8345677 0985132625 7 7 0 3 3 5 6 6723924 1110448043 3 7 8 2 2 8 6 6432118
		Consensus	GACCTTCAGG 8 C 12 13 G 1 7 GGAGAGC
Africa	Namibia	38	A...A....A 6 . 14 12 . . . . . AT..
	Namibia	22	..G..C.TT. . . 11 11 . . . . . A...
	South Africa	32	AGG....T. . . . 11 11 . . . . .
	South Africa	33	.G. . . . . . . 12 . . . . . T..
	South Africa	40	.G....T. . . . . T..
	C. Africa	7	A. . . . . 2 . . . . .
	C. Africa	6	A. . . . . 2 . . . . .
	Zaire	8	.G. . . . . G . 14 . 2 . . . . .
	Zaire	65	A. . . . . TTA . . 11 20 A . . . . . A.
	Zaire	9	A.G.A....A 7 . 14 12 . . . . . AT..
Europe	United Kingdom	26	..G..C.TT. . . 11 11 . . . . . A...
	Germany	61	A.G.A. . . . . 11 9 . . . . . A...
	Germany	64	. . . . . C.TT. . . 11 11 . . . . ATA...
	Germany	62	A. . . . . 11 20 . . . . ATA...
	E. Europe	24	. . . . . C.TT. . . 11 12 . . . . . A...
	Poland/Ukraine	59	..G..C.TT. . . 11 11 . . . . . A...
	S. W. Russia	56	. . . . . C.T. . . . . 2 6 A. . . . .
	Russia	72	A.G.A. . . . . 11 9 . . . . . A...
	Russia	71	. . . . . C.TT. . . 11 11 . . . . . A...
	Turkey	79	A. . . . . 2 . . . . .
Asia	Japan	78	A. . . . . 2 . . . . .
	Japan	76	. . . . . 2 . . . . .
	Japan	77	A. . . . . 2 . . . . .
	China	66	..G. . . . . 2 . . . . .
	China	67	..G. . . . . 13 . . 2 . . . . .
	China	68	. . . . . 13 . . 2 . . . . .
	Cambodia	69	A..G. . . . . 13 . . 2 . . . . .
	Pakistan	57	. . . . . C.TT. . . 11 11 . . . . . A..T
	Siberia	49	. . . . . C.TT. . . 11 12 . . . . . A...
	Siberia	51	. . . . . C.T. . . . 11 . . 2 . . . . .
	Melanesia	10	.GG. . . . . 14 . . . . . T..
Americas	United States	27	A. . . . . T. . . . 8 23 A . . . . . A...
	United States	1	. . . . . 11 11 . . . . . A...
	United States	23	.G. . . . . 14 . . . . . T..
	United States	2	..G. . . . . 2 . . . . .
	United States	4	.GG. . . . . 14 . . . . . T..
	Mexico	17	. . . . . C.TT. . . 11 11 . . . . . A...
	Brazil	12	A.G. . . . . 2 . . . . .
	Brazil	13	..G. . . . . 2 . . . . .
	Brazil	16	. . . . . C.TT. . . 11 11 . . . . . A...
	Brazil	14	A. . . . . 11 . . 2 . . . . .
		Pan	..G....T.A 6 . 11 7 . . 6 .A.A...
		Pongo	..G....T.A . . 10 14 . - 6 .A.A.C.

Numbers below sites indicate number of dinucleotide repeats. Sites and dinucleotides present at each are as follows: 1773 $(CT)_n$, 1808 $(CT)_n$, 1832 $(GT)_n$, 1858 $(GC)_n$, and 1866 $(GT)_n$.

We investigated patterns of linkage disequilibrium by comparing pairs of sites in order along the chromosome; this provides a set of independent comparisons for tests of significance (Lewontin 1995). Sites containing singletons were excluded from this analysis. In intron 7, comparisons were made among four sites (section one, 268, 551; section two, 481, 540). Significant linkage disequilibrium was observed in each of the three sequential comparisons involving these sites (Fisher's exact test, $P < 0.001$ for each, after Bonferroni correction for multiple tests). Sites 268 and 551 in section one of intron 7 are ~10 kb away from sites 481 and 540 in section two of intron 7. The high level of linkage disequilibrium in intron 7 results in two major haplo-

TABLE 4

Nucleotide polymorphism and divergence at *Dmd*

Locus	Length (bp)	Sample size	S	π (SE) (%)	θ (SE) (%)	Tajima's D	Fu and Li's D	Divergence (SE) *Homo-Pan* (%)	Divergence (SE) *Homo-Pongo* (%)
Intron 7	2389	41	9	0.034 (0.028)	0.088 (0.039)	−1.79*	−3.21**	1.63 (0.26)	3.18 (0.37)
Intron 44	3000	41	19	0.141 (0.079)	0.148 (0.056)	−0.16	−0.61	0.90 (0.17)	2.60 (0.30)

Polymorphism and divergence statistics are based on nucleotide substitutions only and do not include insertion/deletion variants. Divergence is based on a single randomly chosen allele from each species. * $P < 0.05$; ** $P < 0.005$.

types (represented by YCC individuals 32 and 8; Table 2). In intron 44, 13 sequential pairwise comparisons were made among 14 sites (1, 191, 681, 814, 834, 1160, 1224, 1353, 1830, 1858, 2374, 2423, 2532, and 2691). Significant linkage disequilibrium was observed in three of these comparisons (Fisher's exact test, $P < 0.0001$ for sites 834–1160 and 1160–1224; $P < 0.05$ for sites 2374–2423, after Bonferroni correction).

None of the 3 comparisons between pairs of sites in intron 7 contained all four gametic types, while 6 of the 13 comparisons between pairs of sites in intron 44 contained all four gametic types. Thus, more recombination is observed among the sequences in intron 44 than in intron 7, consistent with the mapping data in Figure 1. We also calculated the neutral recombination parameter, γ, from the polymorphism data at intron 44 using the method of Hey and Wakeley (1997). This provided an estimate of the per-site population recombination rate, $2Nc_f = 5.58 \times 10^{-3}$, which corresponds to a rate of 27.9 cM/Mb (where c_f is the female recombination rate assuming $N = 10^4$; *e.g.*, Hammer 1995). This value is substantially larger than the estimate of 15 cM/Mb obtained from mapping data (Figure 1), although the variance associated with both of these estimates is large. If population size is closer to 20,000–30,000 (*e.g.*, Nachman *et al.* 1998; Harris and Hey 1999), then the inferred recombination rate from the sequence data (9.3–14.0 cM/Mb) is in better agreement with the estimate from mapping data. γ could not be calculated for either portion of intron 7 because there are no incongruent pairs of sites in these data; the maximum-likelihood estimate of γ in this case is zero (Hey and Wakeley 1997).

The geographic distribution of nucleotide variation at each intron is shown in Table 5. For intron 7, nucleotide diversity is substantially lower in the non-African samples (π ranges from 0 to 0.025%) than in the African sample ($\pi = 0.08\%$). The two major haplotypes at intron 7 are both present in Africa, but only one is present out of Africa. For intron 44, nucleotide diversity in the non-African samples (π ranges from 0.111 to 0.144%) is more than half the value observed in the African sample ($\pi = 0.173\%$). Surprisingly, for both introns, the Asian sample is the least variable and is even slightly, though not significantly, less variable than the sample from the Americas. Average F_{ST} calculated across all populations was six times higher for intron 7 ($F_{ST} = 0.176$) than for intron 44 ($F_{ST} = 0.028$). This overall difference in F_{ST} is attributable to the differences between the two introns in the partitioning of genetic variation between African and non-African populations, as can be seen from the distribution of variation in Tables 2 and 3. Average F_{ST} calculated across all non-African populations was zero for intron 7 and was very small for intron 44 ($F_{ST} = 0.013$).

HKA comparisons involving polymorphism and *Homo-Pan* divergence between intron 7 and intron 44

TABLE 5

Amount and distribution of polymorphisms at *Dmd* introns 7 and 44 by geographic region

Locus	Geographic region	Sample size	*S*	π (SE) (%)	θ (SE) (%)	Tajima's *D*
Intron 7	Africa	10	6	0.080 (0.056)	0.088 (0.053)	−0.409
	Europe	10	1	0.008 (0.012)	0.015 (0.016)	−1.176
	Asia	11	0	0.000 (0.000)	0.000 (0.000)	—
	Americas	10	3	0.025 (0.025)	0.044 (0.032)	−1.572
Intron 44	Africa	10	14	0.173 (0.103)	0.165 (0.085)	0.223
	Europe	10	10	0.144 (0.088)	0.118 (0.064)	0.985
	Asia	11	10	0.111 (0.070)	0.114 (0.060)	−0.105
	Americas	10	9	0.121 (0.076)	0.106 (0.058)	0.617

are shown in Table 6. When all the data are considered, there is only a marginally significant rejection of the null model ($P = 0.08$). However, when the non-African populations are considered collectively, there is a significant reduction in the ratio of polymorphism to divergence at intron 7 relative to intron 44 ($P < 0.05$). A significant reduction is also seen in Europe and in Asia, but not in Africa or the Americas. HKA tests involving *Homo-Pongo* comparisons yield similar results: a significant or marginally significant rejection of the null model is obtained in comparisons involving Asia ($P < 0.05$), Europe ($P = 0.06$), or all non-African populations ($P = 0.08$), but not in comparisons involving the total sample, Africa, or the Americas ($P > 0.10$). We also performed HKA tests comparing *Dmd* intron 7 and *Dmd* intron 44 to another X-linked gene, *Pdha1* (Harris and Hey 1999). The *Pdha1* data consist of 4200 bp surveyed in a worldwide sample of 35 chromosomes. In comparisons using the entire sample, the ratio of polymorphism to divergence is lower at *Dmd* intron 7 than at *Pdha1* (HKA $\chi^2 = 3.41$, $P = 0.06$), but is nearly identical at *Dmd* intron 44 and at *Pdha1* (HKA $\chi^2 = 0.01$, $P > 0.5$).

TABLE 6

HKA tests comparing *Dmd* intron 7 *vs.* intron 44, *Homo vs. Pan*

Geographic region	Locus	*S*	*D*	HKA χ^2	*P* value
Africa	Intron 7	6	39		
	Intron 44	15	27	2.51	NS
Europe	Intron 7	1	39		
	Intron 44	10	27	5.10	<0.05
Asia	Intron 7	0	39		
	Intron 44	10	27	7.01	<0.01
Americas	Intron 7	3	39		
	Intron 44	9	27	2.52	NS
World	Intron 7	9	39		
	Intron 44	19	27	3.08	0.08
non-Africa	Intron 7	4	39		
	Intron 44	15	27	5.01	<0.05

NS, not significant.

DISCUSSION

We investigated the amount and structure of DNA sequence variation at two introns of *Dmd* in a worldwide sample of 41 humans and found that these two introns have strikingly different patterns of genetic variation. In general, intron 44 had a high level of nucleotide diversity, little linkage disequilibrium, no skew in the frequency distribution of polymorphisms, and revealed similar patterns of variation in and out of Africa. Patterns of variation at intron 44 are entirely consistent with a neutral model of molecular evolution. In contrast, intron 7 had a low level of nucleotide diversity, displayed significant linkage disequilibrium extending over 10 kb, a significant excess of rare polymorphisms, and very different patterns of variation in and out of Africa. Jointly, the patterns of variation observed at these two introns are inconsistent with a standard, neutral equilibrium model. The statistical evidence against this model derives from the significantly negative values of Tajima's (1989) *D* and Fu and Li's (1993) *D* for intron 7 (Table 4) and from the significant HKA tests showing reduced variability at intron 7 in non-African populations (Table 6). These patterns are difficult to reconcile with non-equilibrium population-level effects, such as migration or changes in population size, since such effects are expected to affect all loci in a roughly proportional fashion. On the other hand, all of our observations are consistent with positive directional selection acting recently at or near intron 7. Positive directional selection can reduce levels of linked neutral variability, increase levels of linkage disequilibrium, and produce a skew in the frequency distribution toward an excess of rare sites (Maynard Smith and Haigh 1974; Kaplan *et al.* 1989; Tajima 1989; Braverman *et al.* 1995). Moreover, if selection does not act equally in all geographic regions, it may also lead to increased levels of population differentiation (*e.g.*, Stephan 1994; Stephan *et al.* 1998).

The exact nature of selection is difficult to determine from the observed distribution of variation. There are two major haplotypes at intron 7 (represented by YCC individuals 32 and 8) and these haplotypes are three

(YCC 32) and one (YCC 8) mutational steps derived from the ancestral human haplotype, inferred from parsimony using the chimpanzee and orangutan sequences as outgroups. Both of the major haplotypes are present in Africa but only one is present out of Africa. All other haplotypes in our sample are one mutational or recombinational step derived from one of these two major haplotypes. One straightforward explanation for the differing patterns of variation at intron 44 and at intron 7 is a partial selective sweep of the more common haplotype (YCC 32) at intron 7, especially in non-African populations. The fact that variation is reduced primarily in non-African populations suggests that a selective sweep may have occurred concomitant with or following the movement of anatomically modern humans out of Africa. It should be noted that despite the presence of two major alleles, there is no evidence for an excess of variation or for polymorphisms at intermediate frequency as might be expected under prolonged balancing selection (*e.g.*, Hudson *et al.* 1987; Kreitman and Hudson 1991).

The likelihood that selection has acted at or near intron 7 raises the question of which site or sites are the direct targets of selection. The genomic distance in base pairs (d) over which selection is likely to exert a strong effect on levels of linked neutral variability is a function of the strength of selection, s, and the recombination rate per nucleotide, c, and is approximated by $d = (0.01)\, s/c$ (Kaplan *et al.* 1989, p. 896). For example, Wang *et al.* (1999) observed a reduction in neutral variation over a region of only a few kilobases in the vicinity of the 5′ promoter region of the *teosinte-branched 1* locus in maize, a gene that has been a target of strong artificial selection during the domestication of maize. The recombination rate over the entire *Dmd* locus is ~6 cM/Mb, but it may be closer to 4 cM/Mb in intron 7 (Oudet *et al.* 1992; Figure 1), corresponding to $c = 4 \times 10^{-8}$ per nucleotide. Mutations with selection coefficients $<10^{-4}$ are unlikely to have been affected by deterministic processes in ancestral human populations, given most estimates of effective population size (*e.g.*, Hammer 1995; Zietkiewicz *et al.* 1998). If we consider a range of selection coefficients, $0.001 < s < 0.1$, then linked neutral variability is expected to be reduced over a genomic distance ranging from 500 bp to 50 kb. Nucleotide variability at intron 7 is low in both of the segments we sequenced, and these are separated by ~9 kb. This implies that selection coefficients may be $>10^{-3}$, assuming a simple model of genetic hitchhiking (Kaplan *et al.* 1989). Nonetheless, the large size of the window of reduced variation points to the difficulty of identifying the specific site or sites that have been under selection. This stands in contrast to the narrow window of reduced variability in maize at the *tb1* locus (Wang *et al.* 1999) or the narrow window of elevated variability at *Adh* in Drosophila (Kreitman and Hudson 1991). Because average recombination rates in humans (10^{-8} per site) are roughly threefold lower than in Drosophila ($2–3 \times 10^{-8}$ per site; Nachman and Churchill 1996), the size of windows affected by selection on linked sites in humans may, on average, be larger than in flies (assuming equivalent selection coefficients).

The effects of selection are expected to be easiest to detect in genomic regions experiencing low rates of recombination because these regions will contain more potential targets of selection for a given genetic distance. Indeed, our study was motivated by an interest in depicting patterns of variation at a high-recombination gene to capture the distribution of variation that may be closest to neutral, equilibrium values. However, the observed patterns of variation strongly suggest that selection has acted in this region, and these observations raise the possibility that the signature of selection at the molecular level may be common in the human genome. Moreover, the differences in patterns of variation seen at intron 7 and intron 44 highlight that a single functional gene may contain segments with dramatically different evolutionary histories.

Overall, the level of variation we observed in intron 44 is in general agreement with previous surveys of nucleotide variability in this intron (Nachman *et al.* 1998; Zietkiewicz *et al.* 1998). In a sample of 10 alleles surveyed over 1537 bp, Nachman *et al.* (1998) reported nucleotide diversity of 0.187%, a value that is not significantly different from the value reported here. Zietkiewicz *et al.* (1997, 1998) surveyed 7622 bp in a worldwide sample of 250 alleles and reported nucleotide diversity of 0.101%. The slightly lower value obtained by Zietkiewicz *et al.* (1998) may reflect that their survey was based on polymorphism detection using SSCP.

The level of nucleotide variability observed at each intron can be used to estimate the effective population size under the neutral expectation for X-linked genes, $\pi = 3N_e\mu$, assuming a sex ratio of 1. Using the human-chimpanzee divergence values in Table 4, the estimated mutation rates are $\mu = 3.26 \times 10^{-8}$ for intron 7 and $\mu = 1.8 \times 10^{-8}$ for intron 44 assuming a divergence time of 5 mya and a generation time of 20 years. The estimated population sizes are $\sim N_e = 3500$ for intron 7 and $N_e = 26{,}000$ for intron 44. The corresponding coalescence times are ~210,000 years for intron 7 and 1,560,000 years for intron 44. Despite the large variance associated with each of these estimates, these differences underscore the fact that different regions of the genome, and even of the same gene, may provide quite different estimates of parameters that are important for understanding human evolution. Genomic regions that have been influenced by selection at linked sites may provide substantial underestimates of the long-term effective population size for humans. The larger value of N_e obtained from intron 44 is likely to better reflect equilibrium conditions and suggests that a long-term effective population size for humans may be on the order of 30,000 rather than 10,000 (*e.g.*, Hammer 1995).

The geographic patterns reported here are in general agreement with other studies of nucleotide variability in humans in revealing more variation in Africa than in other continental regions (*e.g.*, Harding *et al.* 1997; Zietkiewicz *et al.* 1998; Harris and Hey 1999; Kaessmann *et al.* 1999). This observation is often interpreted as evidence that modern humans throughout the world derived recently from an ancestral African population (*e.g.*, Kaessmann *et al.* 1999), although it has also been pointed out that some of the African diversity may derive from human migration back to Africa (*e.g.*, Harding *et al.* 1997; Hammer *et al.* 1998). One surprising observation in our data is the greater variability in the Americas than in Asia. The Americas are typically thought to have been colonized from Asia, and samples from the Americas typically reveal lower levels of genetic variability than samples from Asia (*e.g.*, Karafet *et al.* 1999). At both intron 7 and intron 44, we observe higher levels of variability in the Americas than in Asia, though in neither case is this difference significant. Surveys of nucleotide variability at additional unlinked loci from multiple populations will be essential for disentangling the effects of population-level processes from selection in shaping variation in different geographic regions.

We thank Mike Hammer for discussions, Isaac Jones for help with sequencing, and Wolfgang Stephan and two anonymous reviewers for helpful comments on the manuscript. This work was supported by the National Science Foundation.

LITERATURE CITED

Abbs, S., R. G. Roberts, C. G. Mathew, D. R. Bentley and M. Bobrow, 1990 Accurate assessment of intragenic recombination frequency within the Duchenne muscular dystrophy gene. Genomics **7:** 602–606.

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.

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

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.

Dvorák, 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.

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

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

Hammer, M. F., T. Karafet, A. Rasanayagam, E. T. Wood, T. K. Altheide *et al.*, 1998 Out of Africa and back again: nested cladistic analysis of human Y chromosome variation. Mol. Biol. Evol. **15:** 427–441.

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.

Harding, R. M., E. Healy, A. J. Ray, N. S. Ellis, N. Flanagan *et al.*, 2000 Evidence for variable selective pressures at Mc1r. Am. J. Hum. Genet. **66:** 1351–1361.

Harris, E. E., and J. Hey, 1999 X chromosome evidence for ancient human histories. Proc. Natl. Acad. Sci. USA **96:** 3320–3324.

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

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

Jaruzelska, J., E. Zietkiewicz, M. Batzer, D. E. C. Cole, J. P. Moisan *et al.*, 1999 Spatial and temporal distribution of the neutral polymorphisms in the last Zfx intron: analysis of haplotype structure and genealogy. Genetics **152:** 1091–1101.

Kaessmann, H., F. Heißig, A. von Haeseler and S. Pääbo, 1999 DNA sequence variation in a non-coding region of low recombination on the human X chromosome. Nat. Genet. **22:** 78–81.

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

Karafet, T. M., S. L. Zegura, O. Posukh, L. Osipova, A. Bergen *et al.*, 1999 Ancestral Asian source(s) of New World Y-chromosome founder haplotypes. Am. J. Hum. Genet. **64:** 817–831.

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.

Kreitman, M., and R. R. Hudson, 1991 Inferring the evolutionary histories of the Adh and Adh-dup loci in Drosophila melanogaster from patterns of polymorphism and divergence. Genetics **127:** 565–582.

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.

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

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., and G. A. Churchill, 1996 Heterogeneity in rates of recombination across the mouse genome. Genetics **142:** 537–548.

Nachman, M. W., V. L. Bauer, S. L. Crowell and C. F. Aquadro, 1998 DNA variability and recombination rates at *X*-linked loci in humans. Genetics **150:** 1133–1141.

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.

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.

Przeworski, M., R. R. Hudson and A. DiRienzo, 2000 Adjusting the focus on human variation. Trends Genet. (in press).

Rana, B. K., D. Hewett-Emmett, L. Jin, B. H. J. Chang, N. Sambuughin *et al.*, 1999 High polymorphism at the human melanocortin 1 receptor locus. Genetics **151:** 1547–1557.

Rieder, M. J., S. L. Taylor, A. G. Clark and D. A. Nickerson, 1999 Sequence variation in the human angiotensin converting enzyme. Nat. Genet. **22:** 59–62.

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.

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 & Hall, New York.

Stephan, W., and C. H. Langley, 1998 DNA polymorphism in *Lyco-*

persicon and crossing-over per physical length. Genetics **150:** 1585–1593.

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

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

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

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

Zietkiewicz, E., V. Yotova, M. Jarnik, M. Korab-Laskowska, K. K. Kidd *et al.*, 1997 Nuclear DNA diversity in worldwide distributed human populations. Gene **205:** 161–171.

Zietkiewicz, E., V. Yotova, M. Jarnik, M. Korab-Laskowska, K. K. Kidd *et al.*, 1998 Genetic structure of the ancestral population of modern humans. J. Mol. Evol. **47:** 146–155.

Communicating editor: W. Stephan