When Did the Human Population Size Start Increasing?

Jeffrey D. Wall* and Molly Przeworski[†,1]

**Department of Ecology and Evolution and [†]Committee on Evolutionary Biology, University of Chicago, Chicago, Illinois 60637*

Manuscript received November 12, 1999
Accepted for publication April 21, 2000

ABSTRACT

We analyze the frequency spectra of all available human nuclear sequence data sets by using a model of constant population size followed by exponential growth. Parameters of growth (more extreme than or) comparable to what has been suggested from mtDNA data can be rejected for 6 out of the 10 largest data sets. When the data are separated into African and non-African samples, a constant size no-growth model can be rejected for 4 out of 8 non-African samples. Long-term growth (*i.e.*, starting 50–100 kya) can be rejected for 2 out of 8 African samples and 5 out of 8 non-African ones. Under more complex demographic models, including a bottleneck or population subdivision, more of the data are compatible with long-term growth. One problem with the data used here is that a subset of loci may reflect the action of natural selection as well as of demography. It remains possible that the correct demographic model is one of constant population size followed by long-term growth but that at several loci the demographic signature has been obscured by balancing or diversifying selection. However, it is not clear that the data at these loci are consistent with a simple model of balancing selection; more complicated selective alternatives cannot be tested unless they are made explicit. An alternative explanation is that population size growth is more recent (*e.g.*, upper Paleolithic) and that some of the loci have experienced recent directional selection. Given the available data, the latter hypothesis seems more likely.

WITH the world's population now in excess of 6 billion, it is clear that the human population size has not remained constant over time. What is still uncertain is when human populations started to expand in size. Did this happen 50–100 thousand years ago (kya) during the Upper Paleolithic (*e.g.*, Rogers and Harpending 1992; Di Rienzo *et al.* 1998; Stiner *et al.* 1999) or only recently after the invention of agriculture roughly 12 kya?

The original arguments for earlier growth were based on mtDNA data (Di Rienzo and Wilson 1991; Rogers and Harpending 1992; Sherry *et al.* 1994). mtDNA data show an excess of rare mutations over equilibrium neutral expectations, which could be a signature of recent population growth (Slatkin and Hudson 1991). Two standard test statistics, *D* (Tajima 1989a) and *D** (Fu and Li 1993), measure whether the observed frequencies of segregating mutations are compatible with the frequencies expected under the standard null model. Some departures from the null model (*e.g.*, recent increases in population size, linkage to a locus under directional selection) lead to an excess of low frequency variants and negative *D* and *D** values, while others (*e.g.*, population subdivision, linkage to a site under balancing selection) tend to cause an excess of intermediate frequency variants and positive *D* and *D** values. The *D* and *D** values for human mtDNA are sharply negative, as are the values for the Y chromosome (Underhill *et al.* 1997; R. Thompson, J. K. Pritchard, P. Shen, P. J. Oefner and M. W. Feldman, unpublished results).

Nonetheless, mtDNA and the nonrecombining portion of the Y chromosome are but two loci experiencing little (Awadalla *et al.* 1999) or no recombination, so the effects of demography are confounded with those of natural selection and genetic drift. Recent positive selection at any one site could produce the observed excess of rare mutations at linked neutral sites (Braverman *et al.* 1995), as could selection against weakly deleterious mutations (for evidence of purifying selection in the mtDNA, see, *e.g.*, Nachman *et al.* 1996; Wise *et al.* 1998). If these loci are indeed affected by selection then it is unclear whether the observed *D* and *D** tell us anything about past human population sizes. The effect of natural selection and other confounding factors can be controlled for if multiple independent nuclear loci are considered: selection affects only a small region while the effects of demography are visible throughout the whole genome. There are now data available from a multitude of microsatellite loci (*e.g.*, Di Rienzo *et al.* 1998; Kimmel *et al.* 1998; Reich and Goldstein 1998), as well as a handful of nuclear sequence studies (*e.g.*, Harding *et al.* 1997; Clark *et al.* 1998; Harris and Hey 1999).

Microsatellite studies generally find evidence for a more ancient start to population growth, but they differ

Corresponding author: Jeffrey D. Wall, 2102 Biological Laboratories, Harvard University, 16 Divinity Ave., Cambridge, MA 02138. E-mail: jdwall@midway.uchicago.edu

[1] *Present address:* Department of Statistics, Oxford University, 1 South Parks Rd., Oxford OX1 3TG, United Kingdom.

on the estimates of the time of expansion and the groups involved. For example, Reich and Goldstein (1998) find evidence for population growth in African populations but not in non-African populations, while Kimmel *et al.* (1998) conclude just the opposite. In addition, the studies consider different scenarios of population growth. Reich and Goldstein (1998) consider a model of sudden population expansion (*i.e.*, a small constant size before some fixed time and a large constant size after this time) and suggest that the lack of a signal of population expansion in non-African populations may have been caused by a bottleneck associated with their initial migration from Africa. In contrast, Kimmel *et al.* (1998) find support for a model of a bottleneck followed by population growth (but not a model of constant size followed by population growth), while Di Rienzo *et al.* (1998) examine the extreme scenario of rapid growth from an initially monomorphic population (*i.e.*, a star phylogeny). In addition to these differences, there is some uncertainty in the estimates of mutation rates and in the underlying mutation process, so the estimates of the time since the start of growth might not be reliable. For example, two recent studies conclude that the onset of growth could have been more recent (*i.e.*, 10–20 kya; Pritchard *et al.* 1999; Gonser *et al.* 2000).

An additional source of information regarding past population sizes comes from nuclear sequence studies. More than half of nuclear loci have positive Tajima's *D* values (Przeworski *et al.* 2000), so they do not provide evidence for recent population growth (*e.g.*, Harding *et al.* 1997; Hey 1997; Zietkiewicz *et al.* 1998). On the basis of three small data sets Hey (1997) has shown that the difference in *D* values between mitochondrial and nuclear loci is larger than expected under either a model of constant population size or a model of recent exponential growth. It remains unknown whether the positive Tajima's *D* values at a larger set of nuclear loci are compatible with the models of population growth proposed from mtDNA and microsatellite data. In this article, we test this question explicitly. We analyze all available studies of human nuclear sequence variation. We focus on whether data are consistent with the predicted effects of population growth and examine whether there are noticeable differences between African and non-African samples. We do not use single nucleotide polymorphism (SNP) data (*e.g.*, Cargill *et al.* 1999; Halushka *et al.* 1999) both because of the high error rate involved in SNP detection when SNPs are not confirmed by sequencing (*e.g.*, Halushka *et al.* 1999) and because of possible biases in the frequency spectrum recovered by variant detection arrays (*cf.* Przeworski *et al.* 2000).

We consider a model of constant population size followed by exponential growth (*cf.* Marjoram and Donnelly 1994) and determine whether the observed *D* and *D** values at nuclear loci are compatible with the distributions of simulated values. We use *D* because of its use in previous studies (*e.g.*, Hey 1997; Fay and Wu 1999) and *D** because simulations suggest that *D** is more effective than *D* in detecting recent population growth (see results). We focus on weak models of growth: a 10-fold or 100-fold increase in size, unlike the more than 150-fold growth of Rogers and Harpending (1992) or the more than 5000-fold estimated increase in census population size (*cf.* Weiss 1984). We do so to be conservative and to maximize the chance that the data are compatible with population growth. We also discuss simulations of more complicated models that include a bottleneck or population subdivision. This discussion is merely qualitative, as distinguishing between more parameter-rich models requires more than the simple analyses performed here (Hey and Harris 1999) and more consistently sampled data.

While clearly of interest, the relation of simple demographic models to debates about human origins is unclear. Indeed, theories of human evolution are often too complex or not specific enough to be testable. The single origin model (*e.g.*, Stringer and Andrews 1988) often assumes explicitly that modern human populations expanded in size as they replaced archaic populations (*e.g.*, Rogers and Harpending 1992). For this reason, evidence for long-term exponential growth from mtDNA and microsatellites has been interpreted as support for the single origin model (Harpending *et al.* 1998). However, the multiregional hypothesis (*e.g.*, Wolpoff *et al.* 1984) is compatible with both an early or a late start for population growth. As this article and others demonstrate, nucleotide polymorphism data can fruitfully be used to test specific demographic models. However, no conclusions can be drawn about more general models of human evolution until these are better specified.

MATERIALS AND METHODS

We examine all human nuclear sequence data sets for which frequency data were available; the sample size (n) was at least 10, and the number of segregating sites (S) was at least four. Some studies were excluded (*e.g.*, Hammer *et al.* 1997; Jin *et al.* 1999; Peterson *et al.* 1999) because of biases in the process of data collection (*e.g.*, polymorphisms not uniformly assayed in all individuals). We also exclude all human major histocompatibility complex (MHC) loci, which are likely to be affected by strong selection (either directly or at linked loci). Only biallelic mutations (both point mutations and indels) were included. The findings are essentially the same when overlapping mutations (*e.g.*, in *Pdha1*, Harris and Hey 1999, or in *Dys44*, Zietkiewicz *et al.* 1998) are included (results not shown). Heterozygosity was calculated as π, the (per-site) average frequency of pairwise differences (Tajima 1983). For *Dys44*, the frequencies of alleles were determined from a larger sample (*cf.* Zietkiewicz *et al.* 1998), since data from the original ascertainment sample were not available.

We assume a neutral infinite-sites model for our simulations. The *P* values for *D* and *D** were determined directly from simulations that first generate genealogies and then place exactly the number of observed segregating sites on the tree

(*cf.* Hudson 1993). A total of 10^5 replicates were run for each parameter combination described below. The simulations require some assumption about the population recombination rate $C = 4Nr$ (N is the effective population size and r is the recombination rate per locus per generation). Assuming no recombination ($C = 0$) is generally conservative for assessing the significance levels of D and D^* (Wall 1999), but this assumption may not be appropriate; most nuclear data sets show evidence of recombination (see discussion). Since on the intragenic scale genetic maps based on pedigree data are not very precise, we estimate recombination rates directly from the patterns in the sequence data. We assume a constant rate of crossing over per base pair and no gene conversion. The estimator C_{HRM} summarizes the data using the estimated minimum number of recombination events (R_M; *cf.* Hudson and Kaplan 1985) and the observed number of distinct haplotypes (H) and returns the value of C that maximizes the likelihood lik($C|H, R_M$) (see Wall 2000 for more details). It is roughly unbiased under a constant population size model, has relatively low mean squared error, and can be calculated for large polymorphism data sets (Wall 2000; J. D. Wall, unpublished results). We incorporate population growth (see below) directly into the null simulations used to estimate C. C_{HRM} could not be calculated for *Dys44*, since haplotype data were not available for that locus, or for *Duffy* and *Dmd7*, because the sequence studied was not contiguous. We assume $C = 0$ for all simulations involving these three loci.

We model either a constant population size or a constant population size followed by exponential growth (*cf.* Marjoram and Donnelly 1994). For the latter, an equilibrium population of size $N = 10^4$ starts at time T to grow at a constant rate to a current population size of 10^5 or 10^6. T, the date of the onset of growth varies from 0 kya to 100 kya, assuming an average generation time of 20 yr.

The recombination rate for each locus is estimated from exponential growth simulations for the whole sample. We run simulations to estimate C_{HRM} for values of T that are multiples of 5 kya and then use linear interpolation to estimate C_{HRM} for other values of T. For *Lpl* and certain values of T, C_{HRM} could not be calculated because the estimated likelihood of the data was 0. This might be because incomplete phase information was available, leading to an underestimate of the number of distinct haplotypes. When this happens, we estimate C solely from the observed R_M [*i.e.*, we take the value of C that maximizes lik ($C|R_M$)]. All simulations use the same growth rates (for a given value of T), except for a simple correction for X-linked loci (which have 3/4 the population size of autosomal loci under the standard neutral model). We consider worldwide samples as well as exclusively African and non-African samples. Most simulations were run using modifications of programs kindly provided by R. R. Hudson.

In addition to simulations of a constant population size followed by exponential growth, for *Lpl* we run simulations of a symmetric island model of geographic subdivision. The model has four islands (meant to correspond loosely to African, European, Asian, and Melanesian populations), and migration rates are taken to correspond roughly to an F_{ST} of 0.15 (*cf.* Takahata 1983; $4Nm = 3.188$, when $N = 10^4$). Actual F_{ST} values between continental populations are often less than this value (Cavalli-Sforza *et al.* 1994), but we have opted to be conservative by maximizing the effect population structure has on the distributions of D and D^*. Each individual in the sample is assigned to one of the four islands on the basis of their ethnicity. The numbers of sampled individuals from each island are 48 (Africa), 94 (Europe), 0 (Asia), and 0 (Melanesia). T is the same for all demes in simulations that include exponential growth.

Additional simulations consider population size reductions ("bottlenecks") followed by exponential growth at time T. Stepwise changes in population sizes are straightforward to implement in a coalescent setting (Tajima 1989b). We consider a model of a constant ancestral population size of $N = 10^4$, followed by a reduction in population size to $N = 10^3$ lasting 10 thousand years (kyr), followed by exponential growth to a current population size of $N = 10^6$. The time since the start of growth varies from $T = 0$–100 kyr, and a generation time of 20 yr is assumed. We present only these limited simulations because our interest is in broad qualitative trends.

RESULTS

Table 1 summarizes some general information about the loci considered. Levels of heterozygosity at the loci studied here are comparable with those reported from previous studies (*e.g.*, Li and Sadler 1991). There is no clear trend in the frequency spectra: 7 out of 12 loci have positive D values, while 4 out of 11 loci have positive D^* values. When the D values of the largest data sets are compared with each other (*cf.* Hey 1997), it is found that *Xq13.3*'s value differs significantly from those of *Dys44*, *Pdha1*, and β-*globin* (results not shown).

To illustrate the effect of recent exponential growth on the distribution of D and D^*, we choose the largest locus with positive D and D^* values (*Lpl*) and the largest locus with negative D and D^* values (*Xq13.3*). For these two loci, we run simulations where the population size is constant at $N = 10^4$, then at time T starts growing exponentially until it reaches $N = 10^6$ at the present. Figures 1 and 2 show the middle 95% of simulated D and D^* values, as a function of T. Figure 1, A and B, shows simulations of D and D^*, respectively, for *Lpl* (assuming $C = 0$), while Figure 2, A and B, illustrates D and D^* for *Xq13.3* (with $C = 0$). The actual values of D and D^* are highlighted for comparison. As T increases, the expected values of D and D^* decrease. Note that the expected value of D^* decreases more rapidly than that of D; this suggests that D^* is more effective for detecting recent increases in population size. Further simulations confirm this (results not shown). For $T \approx$ 50–100 kya, as suggested by Rogers and Harpending (1992), Sherry *et al.* (1994), and others, the observed values of D and D^* for *Lpl* fall outside the 95% confidence interval of simulated values; when realistic recombination rates are assumed, the discrepancy between actual and simulated values is much greater (see below). The values for *Xq13.3* are inside the 95% confidence interval for any T (Figure 2A) or any $T > 10$ kya (Figure 2B) in the interval shown. However, if $T \gg 100$ kya, as in models discussed by Hawks *et al.* (2000), even the D and D^* values for *Xq13.3* are significantly too high (*e.g.*, for $T = 2$ mya and 600-fold growth, $P < 0.01$ for both).

We quantify the effect of T on D and D^* for other loci by determining for which values of T the actual values of D and D^* lie within the middle 95% of simulated D values. Unlike above, these simulations use a recombination rate that is estimated from the data (see

TABLE 1

Polymorphism data for human data sets

Locus	n[a]	Bps	S[b]	π (%)[c]	D*[d]	D[e]	C_{HRM}[f]	Source
Lpl	142	9700	87	0.190	0.926	0.542	115[g]	Clark *et al.* (1998)
Ace	22	24000	78	0.098	−0.286	0.391	17	Rieder *et al.* (1999)
Dys44	250	7622	34	0.093	—[h]	0.744	—[i]	Zietkiewicz *et al.* (1998)
Xq13.3	70	10163	33	0.033	−3.346[j]	−1.616	1.8	Kaessmann *et al.* (1999)
Pdha1	35	4194	24	0.178	0.851	0.973	6.2	Harris and Hey (1999)
Duffy[k]	82	1931	22	0.128	0.188	−0.449	—[l]	Hamblin and Di Rienzo (2000)
Dmd44[m]	41	3000	17	0.136	−0.366	0.089	68	Nachman and Crowell (2000)
β-*globin*	349	2670	19	0.157	−0.593	1.058	21	Harding *et al.* (1997)
Dmd7[l]	41	2389	12	0.051	−2.631[j]	−1.725[j]	—[l]	Nachman and Crowell (2000)
Zfx	336	1089	10	0.082	0.469	−0.938	3.7	Jaruzelska *et al.* (1999)
Mc1r	242	951	6	0.114	−1.063	0.195	0.5	Rana *et al.* (1999)
Hprt	10	2485	4	0.038	−1.127	−1.245	1.0	Nachman *et al.* (1998)

[a] Sample size.
[b] Number of segregating sites.
[c] Average number of pairwise differences per base pair, in percentage (*cf.* Tajima 1983).
[d] From Tajima (1989a).
[e] From Fu and Li (1993).
[f] Sequence-based estimate of the rate of recombination per locus.
[g] These values are probably underestimates since available phase information was incomplete.
[h] Could not be calculated because the data from the original ascertainment sample are unavailable.
[i] Could not be calculated because haplotype information is unavailable.
[j] $P < 0.05$ (two tailed). Coalescent simulations with no recombination were used to assess significance levels.
[k] Includes both *Duffy* and the noncontiguous 5′ region sequenced.
[l] Sequence not contiguous.
[m] This region is in the same intron as *Dys44*.

materials and methods). This is shown in Table 2, for an ancestral population size of $N = 10^4$ and current population sizes of $N = 10^5$ or $N = 10^6$. For 10-fold growth in population size, four loci are inconsistent with exponential growth starting 50 kya. For 100-fold growth, six are inconsistent with $T = 50$ kya and nine with $T = 100$ kya. In contrast, two loci are inconsistent with $T = 0$. (Note that we are not correcting for the use of two test statistics. If we do, the qualitative conclusions are unchanged.)

One of the main conclusions to emerge from studies of human variation is a greater variability in Africa (*e.g.*, Cann *et al.* 1987; Bowcock *et al.* 1994; Halushka *et al.* 1999). To determine whether there is a geographical component to the patterns observed, we partition our data sets into African and non-African samples. The sampling scheme for non-Africans varies greatly, from 34 chromosomes from one population (*Duffy*) to one or few individuals from dozens of populations (*Xq13.3*). All non-African samples include some Europeans; *Dys44*, *Xq13.3*, *Dmd44*, β-*globin*, and *Dmd7* include Asian populations as well. For each, we run the same exponential growth simulations as before. Table 3 summarizes the data sets and shows the results of these simulations for the eight loci that provided geographic information and satisfied minimal size requirements ($n \geq 10$ and $S \geq 4$ in both samples). While our findings confirm the observation of higher levels of polymorphism in African *vs.* non-African populations, other systematic differences seem more difficult to identify. The D values for non-African samples are generally higher than the corresponding D values for African samples (true for 6 out of 8 loci), but perhaps more interesting is that four of the non-African samples (but none of the African samples) have significantly high D values even when there is no growth (*i.e.*, $T = 0$). The P values for these four data sets become vanishingly small under long-term exponential growth (*i.e.*, $T = 50$ kya). In contrast, there seems to be no systematic difference in D* values between African and non-African populations, and one African and one non-African sample show the opposite pattern of significantly negative D values when $T = 0$. Overall, five out of eight non-African and two out of eight African samples are inconsistent with a model of 100-fold growth starting 50 kya.

A model of constant population size followed by exponential growth is probably too simplistic. With the inclusion of additional features, such as a population bottleneck or population subdivision, more data are compatible with an older onset of growth. We highlight this by examining how alternative demographic assumptions affect the distribution of D and D* values for the total *Lpl* data set. Figure 3 shows the middle 95% of simulated D and D* values for a model of a population bottleneck followed by exponential growth, as a function of the time since the end of the bottleneck (see

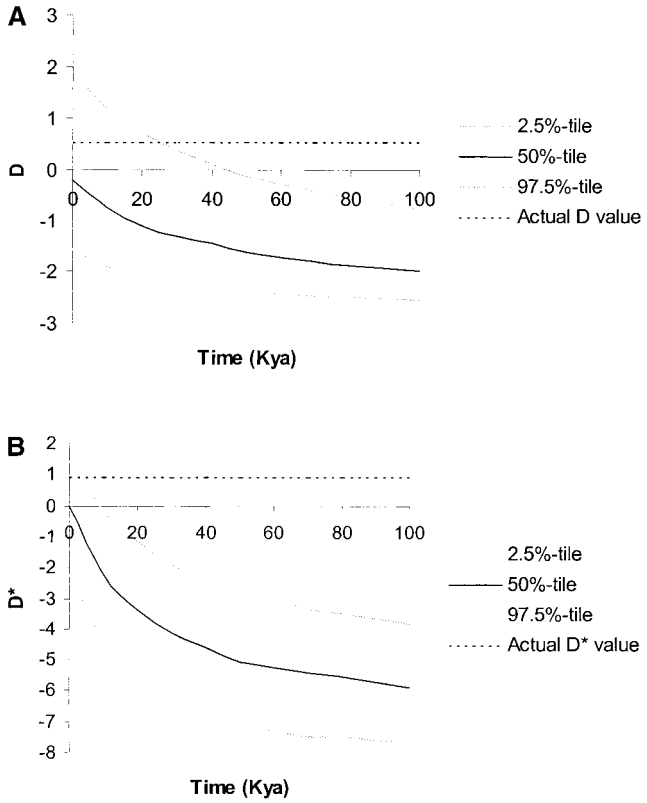

Figure 1.—*D* and *D** *vs. T*, the time since the onset of exponential growth, for *Lpl*. The population size is constant at $N = 10^4$ before *T*, then grows exponentially to a current size of $N = 10^6$. A generation time of 20 yr is assumed. A total of 10^5 simulations are run for $T = 0$–100 kya (increment 10 kyr). A shows simulations for Tajima's *D* ($C = 0$), while B shows simulations for Fu and Li's *D** (with $C = 0$). The simulated 2.5% tile, 50% tile, and 97.5% tile are plotted, as well as the actual values of *D* and *D**.

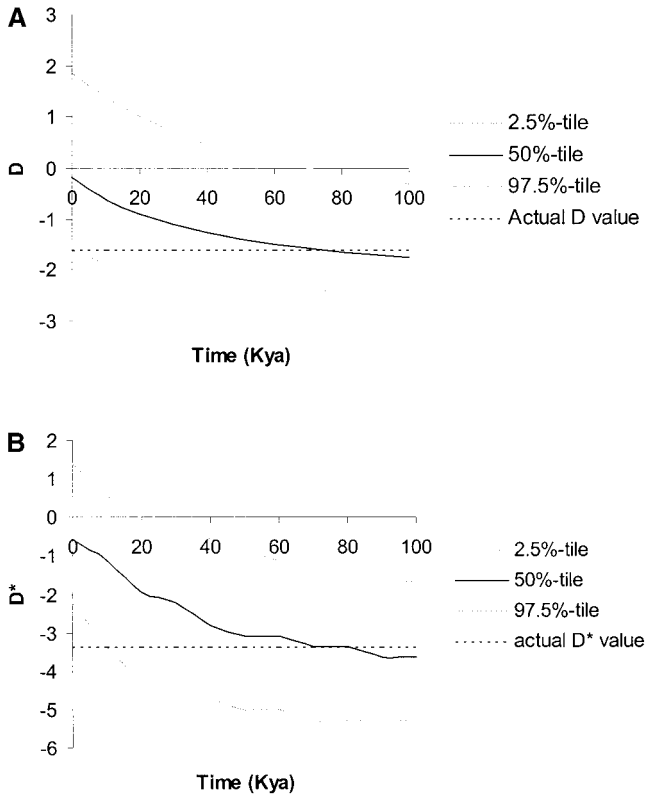

Figure 2.—*D* and *D** *vs. T*, the time since the onset of exponential growth, for *Xq13.3*. The population size is constant at $N = 10^4$ before *T*, then grows exponentially to a current size of $N = 10^6$. A generation time of 20 yr is assumed. A total of 10^5 simulations are run for $T = 0$–100 kya (increment 10 kyr). A shows simulations for Tajima's *D* ($C = 0$), while B shows simulations for Fu and Li's *D** (with $C = 0$). The simulated 2.5% tile, 50% tile, and 97.5% tile are plotted, as well as the actual values of *D* and *D**.

materials and methods). Figure 3A shows simulated Tajima's *D* values, while Figure 3B shows simulated Fu and Li's *D** values. The actual values are highlighted for comparison. The specific parameters used were chosen to maximize the chance that the observed *D* and *D** would be compatible with long-term exponential growth. As can be seen by comparing Figure 1 with Figure 3, the *Lpl* data set is now compatible with an older onset of growth (roughly 46 kya instead of 25 kya for *D* and 7 kya instead of 3 kya for *D**). If recent population growth is assumed, then the effect of a bottleneck before the start of growth decreases as the sample size increases (results not shown). Anatomically modern humans are thought to have reached Australia roughly 50–60 kya (Roberts *et al.* 1994). If $T > 50$ kya, stronger bottlenecks will result in lower values of *D* and *D** (results not shown).

The presence of population structure often leads to higher expected *D* and *D** values. To test the magnitude of this effect, we consider an island model of geographic subdivision (see materials and methods). Figure 4 shows the middle 95% of simulated *D* (Figure 4A) and *D** (Figure 4B) values for *Lpl*. As implemented, geographic subdivision has a relatively minor effect, despite the low migration rate used (see materials and methods); the range of compatibility increases from 0–25 kya in Figure 1A to 0–37 kya in Figure 4A and from 0–3 kya in Figure 1B to 0–6 kya in Figure 4B. If an equilibrium island model with unequal island sizes is used (*cf.* Relethford and Harpending 1995; Relethford and Jorde 1999), the effect on *D* and *D** is almost the same (results not shown).

DISCUSSION

Nuclear sequence data conflict with other genetic loci: In this article we examined the frequency spectrum of segregating mutations at nuclear loci in humans. mtDNA and Y chromosome data show a substantial excess of rare mutations (*i.e.*, *D* and *D** are strongly negative) over equilibrium neutral expectations (Cann *et al.* 1987; Di Rienzo and Wilson 1991; Underhill *et al.* 1997). Researchers have argued that the sharply negative *D* value reflects an expansion in population size that

TABLE 2

Range of *T* values consistent with the observed Tajima's *D* values

	D[a]		*D**[b]	
Data set	10-fold growth[c]	100-fold growth	10-fold	100-fold
Lpl	0–4	0–3	0–2	0–1
Ace	0–100+	0–97	0–100+	0–66
Dys44[d]	0–20	0–11	—[e]	—[e]
Xq13.3	3–100+	2–100+	14–100+	9–100+
Pdha1	0–40	0–23	0–29	0–18
Duffy	0–100+	0–100+	0–100+	0–32
Dmd44	0–100+	0–70	0–100+	0–92
β-*globin*	0–5	0–3	0–100+	0–14
Dmd7	4–100+	3–100+	2–100+	1–100+
Zfx	0–100+	0–100+	0–100+	0–11
Mc1r	0–100+	0–100+	0–100+	0–80
Hprt	0–100+	0–100+	0–100+	0–100+

Range of times since the onset of exponential growth are in kya.

[a] From Tajima (1989a).

[b] From Fu and Li (1993).

[c] Range of *T* values for which the observed *D* falls within the middle 95% of simulated *D* values. Simulations use a value of *C* estimated from the data (see materials and methods).

[d] We take $C = 0$ since haplotype data are unavailable.

[e] *D** could not be calculated because data from the original ascertainment sample are unavailable.

TABLE 3

Comparison of African *vs.* non-African samples

						D		*D**	
Locus	*n*[a]	*S*[b]	π (%)[c]	*D*[d]	*D**[e]	10-fold[f]	100-fold	10-fold[f]	100-fold
Lpl Afr.	48	77	0.198	0.370	1.359	0–18	0–11	None[g]	None[g]
Lpl non-Afr.	94	60	0.173	1.405	1.288	None[g]	None[g]	None[h]	None[h]
Ace Afr.	10	70	0.108	0.228	−0.090	0–100+	0–100+	0–100+	0–100+
Ace non-Afr.	12	44	0.080	1.466	0.912	None[g]	None[g]	0–24	0–19
Dys44 Afr.	115	32	0.099	0.776	—[i]	0–33	0–19	—[i]	—[i]
Dys44 non-Afr.	135	21	0.085	1.950	—[i]	None[h]	None[h]	—[i]	—[i]
Xq13.3 Afr.	23	24	0.035	−1.703	−1.975	7–100+	6–100+	0–100+	0–100+
Xq13.3 non-Afr.	47	17	0.031	−0.556	−1.674	0–100+	0–100+	0–100+	0–100+
Duffy Afr.	48	9	0.052	−0.695	−0.607	0–100+	0–100+	0–100+	0–100+
Duffy non-Afr.	34	15	0.158	0.849	1.141	0–100+	0–70	0–39	0–25
Dmd44 Afr.	10	12	0.159	0.571	0.540	0–100+	0–92	0–100+	0–100+
Dmd44 non-Afr.	31	14	0.125	0.237	0.261	0–100+	0–72	0–100+	0–64
β-*globin* Afr.	103	16	0.098	−0.423	−0.468	0–100+	0–100+	0–100+	0–54
β-*globin* non-Afr.	246	14	0.165	2.251	0.204	None[g]	None[g]	0–100+	0–15
Dmd7 Afr.	10	9	0.133	−0.005	−0.595	0–100+	0–100+	0–100+	0–100+
Dmd7 non-Afr.	31	4	0.011	−1.889	−3.022	14–100+	10–100+	14–100+	10–100+

[a] Sample size.

[b] Number of segregating sites.

[c] Average number of pairwise differences per base pair, in percentage (*cf.* Tajima 1983).

[d] From Tajima (1989a).

[e] From Fu and Li (1993).

[f] Range of *T* values for which the observed *D* falls within the middle 95% of simulated *D* values. Simulations use a value of *C* estimated from the data (see materials and methods).

[g] When $T = 0$ (no growth), the observed *D* has two-tailed $P < 0.01$.

[h] When $T = 0$ (no growth), the observed *D* has two-tailed $P < 0.05$.

[i] *D** could not be calculated because the data from the original ascertainment sample are unavailable.

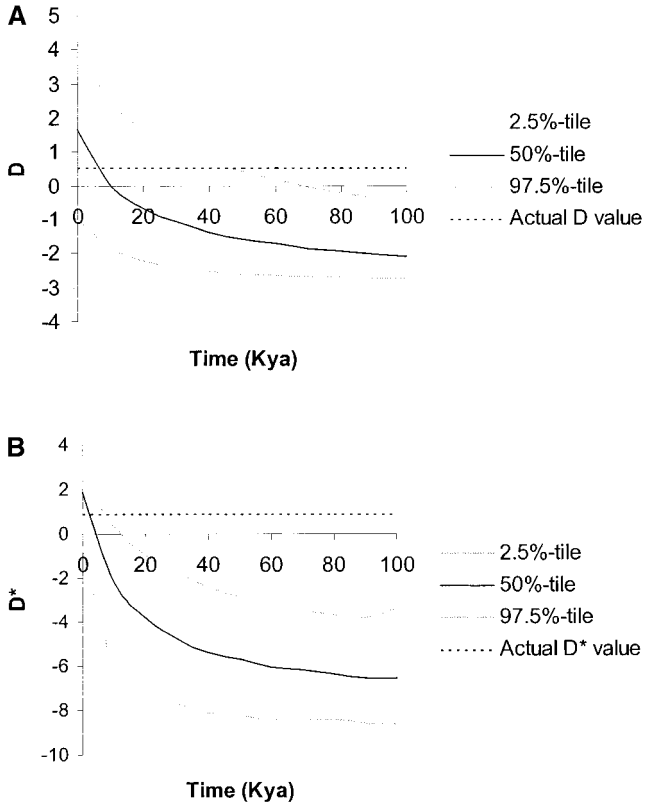

Figure 3.—*D* and *D** *vs. T*, the time since the end of the bottleneck and the onset of exponential growth, for *Lpl*. A population size of $N = 10^4$ is reduced 10-fold for 10,000 years, ending at time *T*. Then the population size increases exponentially to a current size of $N = 10^6$. A generation time of 20 yr is assumed. A total of 10^5 simulations are run for $T = 0$–100 kya (increment 10 kyr). A shows simulations for Tajima's *D* ($C = 0$), while B shows simulations for Fu and Li's *D** (with $C = 0$). We plot the actual values of *D* and *D** as well as the simulated 2.5% tile, 50% tile, and 97.5% tile.

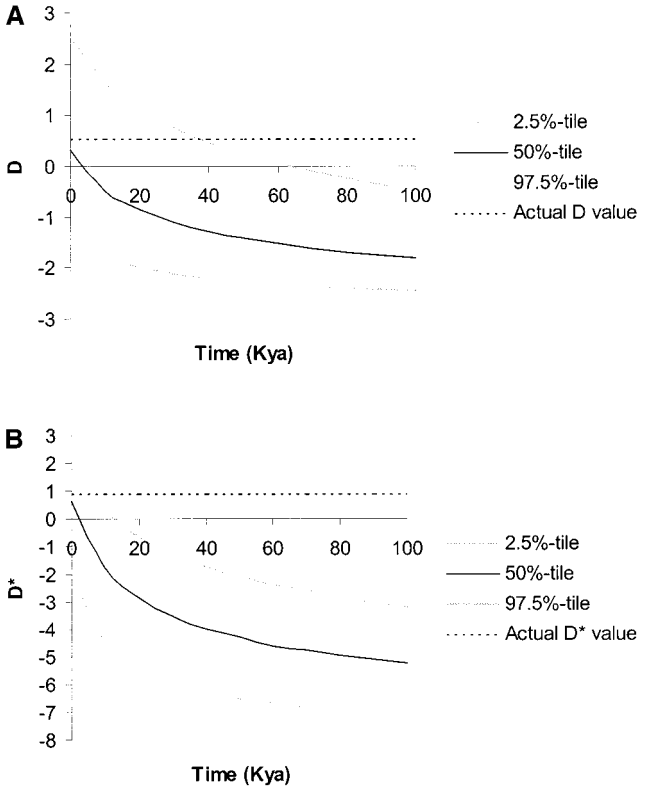

Figure 4.—*D* and *D** *vs. T*, the time since the onset of exponential growth, for *Lpl* with a symmetric island model. The population size is constant at $N = 10^4$ before *T*, then grows exponentially to a current size of $N = 10^6$. A generation time of 20 yr is assumed. A total of 10^5 simulations are run for $T = 0$–100 kya (increment 10 kyr). A shows simulations for Tajima's *D* ($C = 0$), while B shows simulations for Fu and Li's *D** (with $C = 0$).

occurred ~50–100 kya (*e.g.*, Rogers and Harpending 1992; Sherry *et al.* 1994; Rogers 1995). If so, we would expect to observe the effects of this expansion throughout the genome (Hey 1997). However, the nuclear data are not consistent with this scenario, even though our most extreme model of growth (100-fold growth over the past 100 kyr) is still less extreme than has been commonly proposed (*e.g.*, Weiss 1984; Rogers and Harpending 1992). While expansion should lead to predominantly negative *D* and *D** values, Tables 1 and 3 show a roughly equal number of positive and negative values. Since *D* and *D** measure different (but related) aspects of the data, we expect the ranges of compatible *T* values for *D* and *D** to be correlated but not identical in Tables 2 and 3. This is what we observe: For some loci, the *D* and *D** values are large enough such that they are inconsistent with a model of exponential growth starting >50 kya. This is the case for 6 out of 12 loci in Table 2, two out of eight African samples, and five out of eight non-African samples (*cf.* Table 3). Contrary to the claim of Jorde *et al.* (2000), all available data do not support the long-term population growth model first suggested by mtDNA data. The discordance between the data and a long-term growth model is even larger when previously proposed parameter values (*i.e.*, growth rates and time since the start of growth) are used (*e.g.*, Rogers and Harpending 1992; Kruglyak 1999).

Estimating the rate of recombination: One criticism of our conclusions is that significance levels are not necessarily conservative when recombination rates are estimated from the data. However, there is no reason to expect that C_{HRM} consistently overestimates the true recombination rate, and our null simulations explicitly incorporate the model of population growth that is tested. Constant-size coalescent simulations with fixed values of *C* suggest that the median of the distribution of C_{HRM} values is usually less than or equal to the actual value of *C* (results not shown). In addition, some aspects of the data at many loci (in particular, the nonzero and sometimes large values of R_M) are not consistent with low recombination rates (results not shown). We know that recombination is operating throughout the autosomes and X chromosome, and ignoring this fact

might be problematic. In particular, it might not be appropriate for researchers with nuclear sequence data to assume that $C = 0$ and reconstruct a tree, since a *post hoc* pruning of the data (*i.e*, removing sites and individuals that show evidence of recombination) might bias the results. More important, even if we make the unrealistic assumption that $C = 0$, the qualitative results are the same: all of the loci in Tables 2 and 3 that are inconsistent with 100-fold growth and $T = 50$ kya are still inconsistent if we assume no recombination (results not shown).

Possible explanations: Even though nuclear sequence data do not support a simple model of recent population growth, we nonetheless know that a drastic population expansion occurred at least 12 kya with the advent and spread of agriculture. Furthermore, archaeological evidence suggests that human population sizes have expanded over the last 40–50 kyr or more (Klein 1989). So why the discrepancy between expectations and observations? Three nonexclusive possibilities are that our mutational model is incorrect, that our demographic model is incorrect, or that the patterns of variation at some of the loci have been affected (either directly or indirectly) by natural selection.

All of our simulations have assumed an infinite-sites model, and some researchers have recently suggested that multiple mutations at the same site might be frequent for human polymorphism data (Templeton *et al.* 2000). However, this is likely to be a minor concern. For example, for *Lpl* (the data set with the most segregating sites), the expected number of multiple hits at CpG sites is less than two (taking the CpG mutation rate estimated for *Lpl* in Templeton *et al.* 2000). The effect that two multiple hits would have on *D* and *D** is negligible (results not shown). The expected number of multiple hits in smaller data sets is less than what it is for *Lpl* (results not shown).

Although our simplistic demographic model is likely to be incorrect, the relevant question is whether actual human demography differs from our assumptions in ways that would lead to systematically higher *D* and *D** values. We tested two possible alternative models in Figure 3 (population bottleneck) and Figure 4 (population structure). Although both models tend to produce higher *D* and *D** values (and thus greater concordance between our data and a model of recent population growth), neither is a sufficient explanation for all of the loci examined. Some loci (*e.g.*, β-*globin*) still have *D* values that are too high. More generally, the low *D* and *D** values at some loci (*e.g.*, *Xq13.3* and *Dmd7*) and the high values at other loci (*e.g.*, *Lpl* and β-*globin*) are not both consistent with any simple model of human demography (see also Hey 1997). So it seems likely that selection has influenced the patterns of variation at several of the loci studied. One locus in particular (*Duffy*) is known to be influenced by natural selection (which may explain why the African sample has less diversity than the non-African sample; Hamblin and Di Rienzo 2000). Any model of human history should also include claims of how and where natural selection has affected observed genetic variation. Below we examine two main hypotheses for which loci and what types of selection have been operating.

One possibility is that there has indeed been long-term population growth (*e.g.*, $T > 50$ kya). In this case, the excess of rare variants in mtDNA, the Y chromosome, *Xq13.3*, and *Dmd7* reflects demography while the high *D* and *D** values at *Lpl*, *Dys44*, *Pdha1*, and β-*globin* reflect the action of balancing or diversifying selection. The intermediate *D* and *D** values at other loci such as *Ace* or *Dmd44* could then be due to chance or to demographic factors such as population structure or a bottleneck. (But note that these factors are still insufficient to account for the extremely high *D* values at some loci.) Although a simple model of balancing selection (*e.g.*, Hudson and Kaplan 1988) leads to higher *D* and *D** values (Fu 1996), it also predicts a well-defined peak of polymorphism surrounding the selected site. There is neither a putative selected site nor an observable peak of polymorphism at *Lpl*, *Dys44*, *Pdha1*, or β-*globin*. Theory predicts that it takes a substantial amount of time for balancing selection to affect levels of polymorphism or Tajima's *D* values, so young balanced polymorphisms (*e.g.*, malaria resistance at β-*globin*) should have little effect on levels of polymorphism. Note also that there are few if any examples of balanced polymorphisms in any species aside from MHC loci in mammals and S-allele systems in plants. Even the canonical case of *Adh* in *Drosophila melanogaster* may not be a simple balanced polymorphism (Begun *et al.* 1999). Thus, it seems unlikely that balancing selection has led to higher *D* and *D** values in multiple unlinked human nuclear loci. Other selective models, such as local adaptation, might produce higher *D* and *D** values; however, they have not been well characterized theoretically.

An alternative hypothesis is that the positive (and slightly negative) *D* and *D** values reflect demography, while the significantly negative *D* and *D** values for mtDNA, the Y chromosome, *Xq13.3*, and *Dmd7* reflect the recent effects of directional selection. It is an interesting coincidence that three out of four of these are in areas of little or no recombination. The fourth locus, *Dmd7*, shows only an excess of rare variants outside of Africa, so it cannot be taken as support for the simplest model of growth. Kaessmann *et al.* (1999) deliberately chose their region (*Xq13.3*) to be in an area of reduced recombination because data is easier to analyze when recombination can be ignored. They suggest that loci like *Xq13.3* are "ideally suited for unravelling the evolutionary history of the nuclear genome" (Kaessmann *et al.* 1999, p. 79). However, one consequence of their choice of location is that the patterns of variation at *Xq13.3* (as with mtDNA and the nonrecombining region of the Y chromosome) are especially vulnerable to the

effects of selection at linked loci. Positive selection at a linked locus (Smith and Haigh 1974), possibly outside the region examined, would lead to a reduction in heterozygosity and a shift toward negative *D* values at *Xq13.3* (Braverman *et al.* 1995). There are many nearby candidates for selection. A Grail search (http://compbio.ornl.gov) revealed a putative gene (a purinergic receptor) <5 kb away (0.0007 cM) from the region sequenced by Kaessmann and colleagues. Given the available data and the greater prevalence of directional selection compared with balancing selection in Drosophila (Hey 1999), this hypothesis seems more plausible.

Implications for models of human evolution: The single origin model implicitly assumes that modern human populations expanded as they replaced more archaic hominids throughout the Old World (Harpending *et al.* 1998). This expansion presumably happened before the colonization of Australia 50–60 kya (Roberts *et al.* 1994). The challenge for proponents of the single origin model is to formulate a model that includes long-term population growth and that can explain why observed F_{ST} values are low and why there is no trend toward negative *D* and *D** values in nuclear loci. Any claim about the action of selection at certain loci should be made explicit, considering the discussion of balancing selection above. As mentioned before, our results on the timing of recent population expansions do not directly impact the feasibility of the multiregional model. However, it is still unclear whether an ancestral effective population size of 10^4 is consistent with a continuous occupation of most of the Old World (see, *e.g.*, Harpending *et al.* 1998). Perhaps a less polemic goal would be to construct a model that can account for the differences between African and non-African samples. Such a model would need to explain why diversity levels within Africa are consistently higher than outside of Africa, and why the *D* values for non-African samples at some loci are significantly positive. Further work focuses on whether nonequilibrium demographic models are more consistent with human nuclear sequence data.

B. Payseur, C. Sing, and E. Zietkiewicz generously provided unpublished data, and A. Di Rienzo, M. Hamblin, R. Harding, M. Nachman, and J. Pritchard provided preprints of their work. Also, we thank P. Andolfatto, A. Di Rienzo, R. Hudson, M. Nordborg, N. Takahata, and two anonymous reviewers for helpful discussions and comments on earlier versions of this work. Part of this paper was completed when J.D.W. was at the Graduate University for Advanced Studies (Hayama, Japan), supported by the Monbusho Summer Program in Japan. J.D.W. was partially supported by National Institutes of Health grant 5R01H610847.

LITERATURE CITED

Awadalla, P., A. Eyre-Walker and J. M. Smith, 1999 Linkage disequilibrium and recombination in hominid mitochondrial DNA. Science **286:** 2524–2525.

Begun, D. J., A. J. Betancourt, C. H. Langley and W. Stephan, 1999 Is the fast/slow allozyme variation at the *Adh* locus of *Drosophila melanogaster* an ancient balanced polymorphism? Mol. Biol. Evol. **16:** 1816–1819.

Bowcock, A. M., A. Ruiz-Linares, J. Tomfohrde, E. Minch, J. R. Kidd *et al.*, 1994 High resolution of human evolutionary trees with polymorphic microsatellites. Nature **368:** 455–457.

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, P. Sklar, K. Ardlie *et al.*, 1999 Characterization of single-nucleotide polymorphisms in coding regions of human genes. Nat. Genet. **22:** 231–238.

Cavalli-Sforza, L. L., P. Menozzi and A. Piazza, 1994 *The History and Geography of Human Genes.* Princeton University Press, Princeton, NJ.

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.

Di Rienzo, A., and A. C. Wilson, 1991 Branching pattern in the evolutionary tree for human mitochondrial DNA. Proc. Natl. Acad. Sci. USA **88:** 1597–1601.

Di Rienzo, A., P. Donnelly, C. Toomajian, B. Sisk, A. Hill *et al.*, 1998 Heterogeneity of microsatellite mutations within and between loci, and implications for human demographic histories. Genetics **148:** 1269–1284.

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.

Fu, Y.-X., 1996 New statistical tests of neutrality for DNA samples from a population. Genetics **143:** 557–570.

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

Gonser, R., P. Donnelly, G. Nicholson and A. Di Rienzo, 2000 Microsatellite mutations and inferences about human demography. Genetics **154:** 1793–1807.

Halushka, M. K., J. B. Fan, K. Bentley, L. Hsie, N. Shen *et al.*, 1999 Patterns of single-nucleotide polymorphisms in candidate genes for blood-pressure homeostasis. Nat. Genet. **22:** 239–247.

Hamblin, M. T., and A. Di Rienzo, 2000 Detection of the signature of natural selection in humans: evidence from the Duffy blood group locus. Am. J. Hum. Genet. **66:** 1669–1679.

Hammer, M. F., A. B. Spurdle, T. Karafet, M. R. Bonner, E. T. Wood *et al.*, 1997 The geographic distribution of human Y chromosome variation. Genetics **145:** 787–805.

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.

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

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

Hawks, J., K. Hunley, S.-H. Lee and M. Wolpoff, 2000 Population bottlenecks and pleistocene human evolution. Mol. Biol. Evol. **17:** 2–22.

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

Hey, J., 1999 The neutralist, the fly and the selectionist. TREE **14:** 35–38.

Hey, J., and E. Harris, 1999 Population bottlenecks and patterns of human polymorphism. Mol. Biol. Evol. **16:** 1423–1426.

Hudson, R. R., 1993 The how and why of generating gene genealogies, pp. 23–36 in *Mechanisms of Molecular Evolution*, edited by N. Takahata and A. G. Clark. Sinauer, Sunderland, MA.

Hudson, R. R., and N. L. Kaplan, 1985 Statistical properties of the number of recombination events in the history of a sample of DNA sequences. Genetics **111:** 147–164.

Hudson, R. R., and N. L. Kaplan, 1988 The coalescent process in models with selection and recombination. Genetics **120:** 831–840.

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

Jin, L., P. A. Underhill, V. Doctor, R. W. Davis, P. Shen *et al.*

1999 Distribution of haplotypes from a chromosome 21 region distinguishes multiple prehistoric human migrations. Proc. Natl. Acad. Sci. USA **96:** 3796–3800.

Jorde, L. B., W. S. Watkins, M. J. Bamshad, M. E. Dixon, C. E. Ricker *et al.*, 2000 The distribution of human genetic diversity: a comparison of mitochondrial, autosomal and Y-chromosome data. Am. J. Hum. Genet. **66:** 979–988.

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.

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

Klein, R. G., 1989 *The Human Career*. University of Chicago Press, Chicago.

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

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

Marjoram, P., and P. Donnelly, 1994 Pairwise comparisons of mitochondrial DNA sequences in subdivided populations and implications for early human evolution. Genetics **136:** 673–683.

Nachman, M. W., and S. L. Crowell, 2000 Contrasting evolutionary histories of two introns of the Duchenne muscular dystrophy gene, *Dmd*, in humans. Genetics **155**(4) (in press).

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.

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.

Peterson, R. J., D. Goldman and J. C. Long, 1999 Nucleotide sequence diversity in non-coding regions of ALDH2 as revealed by restriction enzyme and SSCP analysis. Hum. Genet. **104:** 177–187.

Pritchard, J. K., M. T. Seielstad, A. Perez-Lezaun and M. W. Feldman, 1999 Population growth of human Y chromosomes: a study of Y chromosome microsatellites. Mol. Biol. Evol. **16:** 1791–1798.

Przeworski, M., R. R. Hudson and A. Di Rienzo, 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.

Reich, D. E., and D. B. Goldstein, 1998 Genetic evidence for a Paleolithic human population expansion in Africa. Proc. Natl. Acad. Sci. USA **95:** 8119–8123.

Relethford, J. H., and H. C. Harpending, 1995 Ancient differences in population size can mimic a recent African origin of modern humans. Curr. Anthro. **36:** 667–674.

Relethford, J. H., and L. B. Jorde, 1999 Genetic evidence for larger African population size during recent human evolution. Am. J. Phys. Anthropol. **108:** 251–260.

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.

Roberts, R. G., R. Jones, N. A. Spooner, M. J. Head, A. S. Murray *et al.*, 1994 The human colonisation of Australia: optical dates of 53,000 and 60,000 years bracket human arrival at Deaf Adder Gorge, Northern Territory. Q. Sci. Rev. **13:** 575–583.

Rogers, A. R., 1995 Genetic evidence for a Pleistocene population explosion. Evolution **49:** 608–615.

Rogers, A. R., and H. Harpending, 1992 Population growth makes waves in the distribution of pairwise genetic differences. Mol. Biol. Evol. **9:** 552–569.

Sherry, S. T., A. R. Rogers, H. Harpending, H. Soodyall, T. Jenkins *et al.*, 1994 Mismatch distributions of mtDNA reveal recent human population expansions. Hum. Biol. **66:** 761–775.

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

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

Stiner, M. C., N. D. Munro, T. A. Surovell, E. Tchernov and O. Bar-Yosef, 1999 Paleolithic population growth pulses evidenced by small animal exploitation. Science **283:** 190–194.

Stringer, C. B., and P. Andrews, 1988 Genetic and fossil evidence for the origin of modern humans. Science **239:** 1263–1268.

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

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

Tajima, F., 1989b The effect of change in population size on DNA polymorphism. Genetics **123:** 597–601.

Takahata, N., 1983 Gene identity and genetic differentiation of populations in the finite island model. Genetics **104:** 497–512.

Templeton, A. R., A. G. Clark, K. M. Weiss, D. A. Nickerson, E. Boerwinkle *et al.*, 2000 Recombinational and mutational hotspots within the human Lipoprotein Lipase gene. Am. J. Hum. Genet. **66:** 69–83.

Underhill, P. A., J. Li, A. A. Lin, S. Qasim Mehdi, T. Jenkins *et al.*, 1997 Delection of numerous Y chromosome biallelic polymorphisms by denaturing high-performance liquid chromotography. Genome Res. **7:** 996–1005.

Wall, J. D., 1999 Recombination and the power of statistical tests of neutrality. Genet. Res. Camb. **74:** 65–79.

Wall, J. D., 2000 A comparison of estimators of the population recombination rate. Mol. Biol. Evol. **17:** 156–163.

Weiss, K. M., 1984 On the number of members of the genus *Homo* who have ever lived, and some implications. Hum. Biol. **56:** 637–649.

Wise, C. A., M. Sraml and S. Easteal, 1998 Departure from neutrality at the mitochondrial NADH dehydrogenase subunit 2 gene in humans, but not in chimpanzees. Genetics **148:** 409–421.

Wolpoff, M. H., X. Wu and A. G. Thorne, 1984 Modern *Homo sapiens* origins: a general theory of hominid evolution involving the fossil evidence from East Asia, pp. 411–483 in *The Origins of Modern Humans: A World Survey of the Fossil Evidence*. Liss, New York.

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: N. Takahata