Human Population Structure and Its Effects on Sampling Y Chromosome Sequence Variation

Michael F. Hammer,*,†,[1] Felisa Blackmer,* Dan Garrigan,‡ Michael W. Nachman† and Jason A. Wilder*

*Genomic Analysis and Technology Core, Division of Biotechnology and †Department of Ecology and Evolutionary Biology, University of Arizona, Tucson, Arizona 85721 and ‡Department of Biology, Arizona State University, Tempe, Arizona 85287

Manuscript received October 21, 2002
Accepted for publication April 10, 2003

ABSTRACT

The excess of rare variants in global sequencing studies of the nonrecombining portion of the Y chromosome (NRY) has been interpreted as evidence for the effects of human demographic expansion. However, many NRY polymorphisms are geographically localized and the effect of different geographical sampling on patterns of NRY variation is unknown. We use two sampling designs to detect population structure and its effects on patterns of human NRY polymorphism. First, we sequence 26.5 kb of noncoding Y chromosome DNA from 92 globally distributed males representing 35 populations. We find that the number of polymorphisms with singleton variants is positively correlated with the number of populations sampled and that there is a significant negative correlation of Tajima's *D* (TD) and Fu and Li's *D* (FD) statistics with the number of pooled populations. We then sequence the same region in a total of 73 males sampled from 3 distinct populations and find that TD and FD values for the 3 pooled and individual population samples were much less negative than those in the aforementioned global sample. Coalescent simulations show that a simple splitting model of population structure, with no changes in population size, is sufficient to produce the negative values of TD seen in our pooled samples. These empirical and simulation results suggest that observed levels of NRY population structure may lead to an upward bias in the number of singleton variants in global surveys and call into question inferences of population expansion based on global sampling strategies.

PATTERNS of genetic variation within and among human populations contain information about the origin and demographic history of our species. The bulk of the evidence from nuclear and mitochondrial DNA studies has been claimed to support a recent African origin of anatomically modern humans (Cann *et al.* 1987; Vigilant *et al.* 1991; Batzer *et al.* 1994; Cavalli-Sforza *et al.* 1994; Tishkoff *et al.* 1996; Hammer *et al.* 1998; Underhill *et al.* 2001). However, contrasting views of human demography have emerged from analyses of different components of the genome. Early work on mtDNA suggested that human populations expanded in size from a small initial population (Di Rienzo and Wilson 1991; Rogers and Harpending 1992; Harpending 1994). The mtDNA data exhibited an excess of rare mutations over neutral equilibrium expectations, which could be a signature of recent population growth (Slatkin and Hudson 1991). Alternatively, this pattern could result from the action of natural selection (*e.g.*, linkage to a site under directional selection or to sites under weak purifying selection; Tajima 1989a; Fu and Li 1993; Braverman *et al.* 1995).

Two standard test statistics, Tajima's *D* (TD; Tajima 1989a) and Fu and Li's *D** (FD; Fu and Li 1993), measure whether the observed frequencies of segregating mutations are compatible with the frequencies expected under the standard neutral model. Population growth, directional selection, or the presence of weakly deleterious mutations may lead to an excess of low-frequency variants and negative TD and FD values, while population contraction and balancing selection may cause an excess of intermediate-frequency variants and positive TD and FD values (Tajima 1989a; Fu and Li 1993). As the number of published human nuclear DNA sequencing data sets grows, a consensus frequency distribution pattern has not yet emerged. Many nuclear loci show TD values that are positive or close to zero and, hence, do not provide support for a population expansion from a small initial size (Harding *et al.* 1997; Hey 1997; Zietkiewicz *et al.* 1998; Harris and Hey 1999; Nachman and Crowell 2000; Przeworski *et al.* 2000; Koda *et al.* 2001; Martinez-Arias *et al.* 2001; Nachman 2001; Pluzhnikov *et al.* 2002). However, negative TD values at other loci have been interpreted to support the mtDNA-based population expansion hypothesis (Kaessmann *et al.* 1999; Shen *et al.* 2000; Alonso and Armour 2001;

This article is dedicated to the memory of David C. Rowe.

[1]*Corresponding author:* Biosciences West Bldg., Department of Ecology and Evolutionary Biology, University of Arizona, Tucson, AZ 85721. E-mail: mhammer@u.arizona.edu

Stephens *et al.* 2001; Thorstenson *et al.* 2001; Yu *et al.* 2001). Fay and Wu (1999) have shown that simple changes in population size may lead to radically different frequency spectra for markers with different effective population sizes, such as mtDNA and autosomes. Nevertheless, it has been suggested that there is too much heterogeneity in frequency spectra among loci to be compatible with either the standard neutral model or a simple long-term exponential growth model (Harpending and Rogers 2000; Przeworski *et al.* 2000; Wall and Przeworski 2000).

One of the problems associated with interpreting these patterns of human DNA variability concerns the sampling strategies employed by different investigators. Current sampling schemes vary between two extremes: global sampling, in which a small number of individuals from many different populations are used in sequencing surveys, and population-based sampling, in which larger numbers of individuals from fewer populations are sequenced. It is important to ask to what degree different sampling designs lead to increased variance in observed patterns among loci. Recently, Ptak and Przeworski (2002) found evidence that summaries of the allele frequency spectra at a large number of autosomal and X-linked loci are affected by sampling design. For example, they found that surveys that sample few individuals from many localities recover more rare alleles than those that sample many individuals from few localities. Despite the intense interest in the nonrecombining portion of the Y chromosome (NRY) for inferring human demographic history, there have been no analyses of the effect of sampling design on observed patterns of variation. The NRY has the lowest level of polymorphism of the 24 human chromosomes (International SNP Map Working Group 2001). Estimates based on sequencing studies show that there is an average of one nucleotide difference per 10,000 bp between two randomly drawn NRY sequences (Whitfield *et al.* 1995; Shen *et al.* 2000). This estimate is an order of magnitude lower than the genome average of approximately one single nucleotide polymorphism (SNP) for every kilobase (Cargill *et al.* 1999; Halushka *et al.* 1999; Stephens *et al.* 2001; Shen *et al.* 2002). Because most surveys of variation on the NRY were originally motivated by the desire to identify informative SNPs for phylogeographic studies, a global sampling strategy was adopted. Typically, several kilobases of DNA were sequenced in a small sample of males (50–70) from many (30–40) globally distributed populations (Shen *et al.* 2000; Hammer *et al.* 2001; Underhill *et al.* 2001). Thus, each population may be represented by only one or two individuals. The statistically significant excess of rare polymorphisms (and negative TD values) observed in these surveys has been interpreted strictly in terms of a panmixia model with exponential population growth (Harpending *et al.* 1998; Shen *et al.* 2000; Thomson *et al.* 2000). One problem with this interpretation is that most NRY polymorphisms tend to be geographically localized (Jobling and Tyler-Smith 1995; Seielstad *et al.* 1998; Hammer *et al.* 2001) and polymorphisms that occur as singletons in global studies are often later found at intermediate frequencies in SNP genotyping surveys of larger samples from individual populations (Hammer *et al.* 2001). Furthermore, the NRY exhibits one of the highest levels of among-group variance and population structure of any known human genetic system (Hammer *et al.* 2001; Romualdi *et al.* 2002). For this reason, caution is required when interpreting patterns of NRY variation on the basis of the assumption of panmixia.

Our purpose here is to examine sequence variation on the NRY using both global and population-based sampling designs to assess the effects of population structure and/or population growth on patterns of NRY diversity. First, we sequence 26.5 kb of noncoding DNA from the Y chromosomes of 92 globally distributed males, and then we sequence the same region from a sample of 73 males drawn from three distinct populations: the Khoisan from Africa ($n = 25$), the Khalkhs from Mongolia ($n = 24$), and Papua New Guinean highlanders ($n = 24$). The results are consistent with substantial levels of population structure and suggest that global sampling designs may bias NRY sequence surveys toward an excess of polymorphisms with singleton variants.

SUBJECTS AND METHODS

Subjects: DNA sequences were screened for polymorphism in two sampling panels. Panel A included a sample of 92 human males, including 28 from Africa (10 Khoisan, 6 East Bantu, 5 West Bantu, 3 Mbuti, 2 Biaka, 1 Ethiopian, and 1 Egyptian), 13 from the Americas (3 Southwest Amerinds, 3 Surui, 2 Karatiana, 2 Mayans, 1 Navajo, 1 Porch Creek, and 1 Tohono O'Odham), 20 from Europe/Middle East (4 Russians, 5 Ashkenazi Jews, 4 Adygeans, 4 Germans, 1 Turk, 1 Englishman, and 1 Yemenite Jew), 23 from Asia (5 Yakuts, 4 Japanese, 5 Han Chinese, 3 Pakistanis, 1 Khmer Cambodian, 1 Buryat, 1 Forest Nentsi, 1 Khant, 1 Selkup, and 1 Sinhalese), and 8 from Oceania (4 Papua New Guineans, 2 Australian Aboriginal People, and 2 Nasioi) (Figure 1). Seventy-two of these samples were from the Y Chromosome Consortium (YCC) cell line repository (Y Chromosome Consortium 2002). Panel B included 25 Khoisan, 24 Mongolians, and 24 Papua New Guineans. Some of the individuals in the second panel (*i.e.*, 10 Khoisan and 4 Papua New Guineans) were also part of the first panel. We also analyzed two additional data sets resulting from previously published mutation detection experiments (Hammer *et al.* 1998, 2001). The first data set resulted from screening a global sample of 20 males (9 sub-Saharan Africans, 3 Asians, 3 Native Americans, 3 Europeans, and 2 Oceanians; Hammer *et al.* 1998), and the

second was based on a global panel of 57 males (17 sub-Saharan Africans, 15 Asians, 11 Europeans, 7 Native Americans, and 7 Oceanians; Hammer *et al.* 2001). All sampling protocols were approved by the Human Subjects Committee at the University of Arizona.

PCR amplification and mutation detection: Panels A and B were screened for NRY polymorphism using two different approaches. Denaturing high-performance liquid chromatography (DHPLC) was used to screen for polymorphisms in the following regions on the NRY: (1) 11.3 kb of the arylsulfatase D pseudogene (ARSDP; GenBank accession no. AC002992), (2) two Yα5 Alu elements within the 16E4 and 486,O,2 clones (GenBank accession nos. AC003094 and AC002531, respectively), and (3) four noncoding regions originally identified in anonymous clones used as probes to detect restriction fragment length polymorphism variation on the NRY of humans and great apes (Allen and Ostrer 1994). These regions (referred to as anonymous clone regions) include 1.3 kb within clone 4-1 (*DYS188*), 1.1 kb in clone 3-1 (*DYS189*), 1.1 kb in clone 3-11 (*DYS190*), and 1.5 kb in clone 3-8 (*DYS194*). *DYS194* is present on the NRY in four copies and *DYS190* is duplicated. A total of ~10.6 kb of anonymous clone region sequence was analyzed. Sequence information from these regions was used to design primers to amplify shorter fragments for DHPLC analysis. Internal primers (available from authors upon request) were used to generate overlapping products for DHPLC analysis and for automated DNA sequencing.

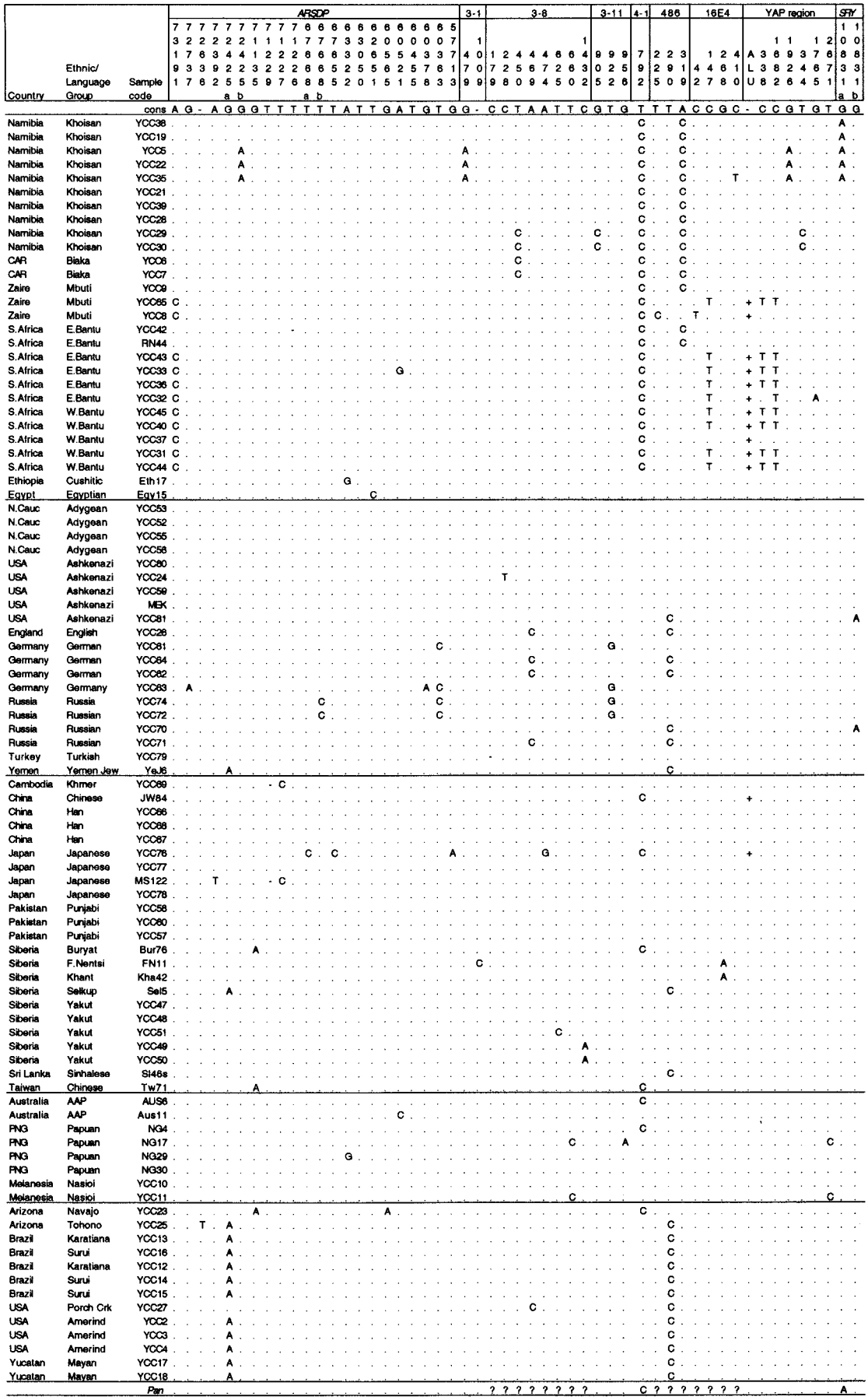
All PCR products producing chromatograms with profiles differing from those of the homoduplex controls were subjected to DNA sequencing. DNA sequencing was performed by standard procedures to identify mutations that altered mobility in DHPLC chromatograms. We were extremely conservative in choosing products for DNA sequencing so as to identify and confirm all possible polymorphisms. The entire 11.3-kb ARSDP and 10.6-kb anonymous clone regions were sequenced from several individuals to assess the error rate in DHPLC. No additional polymorphisms were discovered by sequencing that were not found by DHPLC. The following regions were amplified and subjected directly to DNA sequencing: (1) 2.7 kb of the YAP region (Hammer 1995) and (2) 941 bp of the region upstream of the SRY gene (Whitfield *et al.* 1995). Contiguous sequence was assembled for each individual and aligned using the computer program SEQUENCHER (Gene Codes, Ann Arbor, MI).

Data analysis: Nucleotide diversity, π (Nei and Li 1979), and the proportion of segregating sites, θ (Watterson 1975), were calculated using the program DNASP (Rozas and Rozas 1999). Under mutation-drift equilibrium, both π and θ estimate the neutral parameter $2N_e\mu$ for the NRY, where N_e is the male effective population size and μ is the neutral mutation rate per nucleotide site. Tajima's D (Tajima 1989a) and Fu and Li's D (Fu and Li 1993) were calculated to test the observed mutation frequency distribution for deviations from neutral expectations. Analysis of molecular variance (AMOVA) taking into account the number of mutational differences among haplogroups and exact tests of population differentiation (Raymond and Rousset 1995) were carried out using the ARLEQUIN computer application (Schneider *et al.* 2000).

Computer simulations: To distinguish the expected contributions of population structure, growth, and sampling design on patterns of NRY variation, we simulated samples according to three different models of the neutral coalescent process. To begin, we examined the effect on TD of the sample size taken from an exponentially growing panmictic population, for sample sizes up to 600 chromosomes. Second, we examined the effect on TD of the number of demes sampled from a finite island model of population structure (Maruyama 1970), both with and without population growth. Lastly, we employed a model of population structure in which pairs of populations share common ancestry in a hierarchical fashion, with no migration between demes. This last model will be called the population bifurcation model. We also examined the effect on TD of the number of demes sampled from the population bifurcation model, both with and without growth.

For each model of population structure, the number of sampled demes varied between 5 and 35. No unsampled demes were included in the model. For each of these sampling schemes, an additional level of sampling effect was examined, which includes (1) sampling only 2 haploid individuals from each deme and (2) sampling 20 haploid individuals per deme. A two-phase model of population growth was implemented by assuming that the onset of growth occurred 10^3 generations in the past ($t_g = 10^3$). Before time t_g, the population is assumed to be stationary in size, at $N = 10^3$ (N is the effective number of haploid individuals). Then, at time t_g, the population grows exponentially to $N = 10^5$ in the current generation. Gene genealogies were constructed according to the coalescent probabilities given by Slatkin and Hudson (1991) for a growing population. A total of 40 mutations were then added to each resulting genealogy, according to an infinite sites model.

For the island model of population structure, both strong ($Nm = 1.0$, m is the rate of migration per deme per generation) and weak ($Nm = 10^{-3}$) migration were considered. In this implementation of the island model, m is held constant each generation and migration occurs symmetrically between all demes (Hudson 1990). To parameterize the population bifurcation model, we adopted the Bayesian approach of Wilson *et al.* (2003). Under this model, the first population split occurs τ generations in the past. The prior probability for this time was characterized by a gamma distribution with


ARSDP
3-1
3-8
3-11
4-1
486
16E4
YAP region
SRY
Country
Ethnic/ Language Group
Sample code
cons

$E(\tau) = 5 \times 10^3$ generations ago. The prior distribution of subsequent population splitting times is jointly uniform over the interval $(0, \tau)$. Additionally, each deme was assumed to constitute an equal proportion of the total population size and these proportions were described by a Dirichlet prior distribution. Each simulation bout consisted of 1000 replicates.

RESULTS

Nucleotide diversity in global samples: A total of 26.5 kb of NRY DNA was screened for polymorphism in a sample of 92 globally distributed Y chromosomes. The screened regions—11.3 kb of the ARSDP, 2.7 kb of the YAP region (Hammer 1995), 941 bp upstream of the SRY gene (Whitfield *et al.* 1995), 994 bp encompassing two Yα5 Alu elements, and 10.5 kb of anonymous DNA (Allen and Ostrer 1994)—were all noncoding. A total of 49 polymorphic sites were within the 26.5-kb noncoding region (Figure 1). Three cases of mutational homoplasy were detected because recurrent mutations were found on different NRY haplogroup backgrounds. This brought the total number of mutations to 52: 46 were single nucleotide substitutions (SNPs), 5 were insertion/deletion of a single nucleotide (indels), and 1 was an insertion of an Alu element (YAP; Hammer 1994).

Nucleotide diversity ($\pi \pm$ SD) values varied slightly across the five noncoding regions (ARSDP, 0.009 ± 0.001%; YAP, 0.018 ± 0.004%; SRY, 0.015 ± 0.005%; anonymous DNA, 0.010 ± 0.001%), with the highest π value observed for the Yα5 Alu elements (0.088 ± 0.009%). Nucleotide diversity for the entire 26.5-kb region was 0.014 ± 0.001%, a value very similar to estimates based on other global NRY polymorphism surveys (Table 1). This suggests that the methods used to detect NRY polymorphisms (DHPLC and direct DNA sequencing of PCR products) are comparable. For example, if DHPLC were actually less efficient at detecting polymorphism, we would expect studies based on this method to yield a lower proportion of singletons. In fact, the studies that used DNA sequencing did not find a higher proportion of polymorphisms with singleton variants than the studies based on DHPLC. This supports earlier studies demonstrating the high sensitivity and low error rate of DHPLC (O'Donovan *et al.* 1998).

Levels of nucleotide variability were generally higher in African than in non-African populations (Table 2), consistent with other studies of NRY SNP variation (Hammer *et al.* 1997, 1998; Underhill *et al.* 1997; Shen *et al.* 2000; Underhill *et al.* 2001). When considering π values, Africans were about two to four times more diverse than non-Africans. This discrepancy was not as apparent when considering θ: non-Africans as a whole had higher θ values than Africans; however, no single non-African continental region had a higher θ than Africans (Table 2).

Patterns of nucleotide diversity in global samples: The 26.5-kb noncoding region examined here exhibited statistically significant negative TD and FD values ($P < 0.05$) for the sample of 92 Y chromosomes (Table 1), reflecting a more than twofold excess of singletons over neutral expectations (Figure 2A). A similar observation was made in previous studies of NRY variation using a global sampling strategy (Underhill *et al.* 1997; Shen *et al.* 2000), although earlier studies based on fewer numbers of samples resulted in TD values closer to zero (Hammer 1995; Whitfield *et al.* 1995). We noted a statistically significant negative correlation between TD values in Table 1 and the sample size of each study ($r = -0.796$, $r^2 = 0.634$, $P = 0.010$). A slightly stronger negative correlation resulted when we considered the relationship between TD and the number of "populations" sampled ($r = -0.843$, $r^2 = 0.711$, $P = 0.004$). No significant correlation was observed between TD and the length of the region surveyed ($P = 0.40$).

To further address the relationship between the frequency distribution and the number of populations surveyed, we measured variation within each continental region separately for the 92 Y chromosomes in this survey (Table 2), as well as in the published survey of Shen *et al.* (2000). We calculated TD and FD for the total sample, for combined non-Africans, and for each continental region (*i.e.*, Africans, Asians, Europeans, Oceanians, and Native Americans). For each of these five NRY regions (26.5 kb, SMCY, DFFRY, DBY, and UTY1), a negative correlation between TD or FD and the number of populations sampled was observed. For the 26.5-kb region reported here (Figure 3, A and B), as well as for three of the four genes reported by Shen *et al.* (2000) (excluding DFFRY), FD produced a more significant negative correlation than did TD (data not shown). For three of the five regions (26.5-kb noncoding, SMCY, and DBY), there was a statistically significant negative correlation between FD and the number of populations analyzed (*e.g.*, $P \leq 0.01$), while the correlations for the other two regions (DFFRY and UTY) were marginally

Figure 1.—Polymorphic sites in 26.5 kb of noncoding DNA on 92 Y chromosomes from Africa, Europe/Middle East, Asia, Oceania, and the Americas. Numbering at the top of the figure indicates the position of each polymorphic site in reference to the following sequences: (1) arylsulfatase pseudogene (*ARSDP*); (2) anonymous (Allen and Ostrer 1994) clones 4-1 (*DYS188*), 3-1 (*DYS189*), 3-11 (*DYS190*), and 3-8 (*DYS194*); (3) Yα5 Alu elements within the 16E4 and 486,O,2 clones, respectively; and (4) 2.6-kb YAP sequence (Hammer 1995), SRY gene region (Whitfield *et al.* 1995). Nucleotide positions at the two Yα5 Alu sequences are defined in reference to base positions after the 3′ end of sequencing primers designed to the clone sequences AC003094 and AC002531 (see Hammer *et al.* 2001).

TABLE 1

Nucleotide polymorphism on the NRY

NRY region	Length (bp)	*n*	No. of pops.[a]	*S*	π (%)	Θ (%)	Tajima's *D*	*P*	Fu and Li's *D*	*P*	Reference
26.5 kb	26,500	92	35	46	0.014	0.034	−1.920	<0.05	−2.538	<0.05	This study
SRY	18,300	5	5	3	0.008	0.008	−0.145	>0.10	−0.505	>0.10	Whitfield *et al.* (1995)
YAP	2,638	16	8	3	0.037	0.028	0.250	>0.10	−0.040	>0.10	Hammer (1995)
SMCY	35,311	53	33	44	0.008	0.025	−2.278	<0.01	−4.189	<0.02	Shen *et al.* (2000)
DFFRY	16,669	70	41	17	0.008	0.021	−1.890	<0.05	−1.569	>0.05	Shen *et al.* (2000)
DBY	9,000	70	41	14	0.008	0.032	−2.150	<0.05	−4.234	<0.02	Shen *et al.* (2000)
UTY1	15,317	72	41	20	0.013	0.027	−1.570	>0.05	−2.799	<0.05	Shen *et al.* (2000)
SSCP	5,128	20	12	7	0.022	0.039	−1.429	>0.10	−1.204	>0.10	Hammer *et al.* (1998)
DHPLC	13,444	58	31	16	0.011	0.026	−1.722	>0.05	−2.503	<0.05	Hammer *et al.* (2001)

n, number of chromosomes; *S*, number of segregating sites.
[a] Number of populations or ethnic groups.

statistically significant (*e.g.*, $P = 0.05$ and $P = 0.06$, respectively; data not shown). When all five regions were combined (Figure 4), the correlation between FD and number of populations sampled was highly statistically significant ($r = -0.756$, $r^2 = 0.577$, $P < 0.0001$). A nonparametric Spearman rank correlation test also found a statistically significant monotonic decrease between FD and the number of populations sampled ($r = -0.8005$, $T = -7.56$, $P < 0.0001$). Despite these clear patterns, it is important to point out that not all the data points in Figures 3 and 4 are independent since some of the small samples are subsets of the larger samples.

Patterns of nucleotide diversity in population samples: To discern the effects of sampling on patterns of human NRY nucleotide diversity, we sequenced the same 26.5-kb noncoding region on the Y chromosomes of 73 males representing three distinct populations: the Khoisan of Namibia ($n = 25$), Papua New Guinean highlanders ($n = 24$), and Mongolian Khalks ($n = 24$). In the total sample, we identified 30 SNPs and 3 indels (Figure 5). The nucleotide diversity in this sample ($\pi = 0.014 \pm 0.001\%$) was almost identical to the π for the global survey of 92 chromosomes. However, singletons ($s = 10$) made up a smaller proportion (33.3%) of the segregating sites within populations than within the global sample ($s = 24$ or 47.1%). This was reflected in less negative and statistically nonsignificant TD and FD values in the combined sample of 73 chromosomes ($P > 0.10$; Table 3). The three populations each exhibited different frequency distributions (Figure 2, B–D). The Khoisan had no singletons and several intermediate-frequency sites, while 7 of 11 of the Mongolian polymorphisms were singletons and only 2 were at intermediate frequency. The Papua New Guinean (PNG) frequency distribution was characterized by a lower percentage of singletons (4 of 9 polymorphisms), 2 intermediate-frequency sites, and 1 high-frequency polymorphism. All TD values were slightly or moderately negative, ranging from −0.007 ($P > 0.10$) in the Khoisan to −0.701 ($P > 0.10$) in the Mongolians, but none was statistically significant (Table 3). In contrast, the FD value for the Khoisan was positive and statistically significant (FD = 1.51, $P < 0.02$), signifying a deficiency of singletons.

Population structure: In the population-based sample, only 4 of the 23 nonsingleton polymorphisms and a single NRY haplogroup were shared among the populations (Figure 5). Two polymorphisms were shared among all three populations, 1 SNP was shared between PNG and Mongolians, and 1 indel (YAP) was shared between the Khoisan and Mongolians. While this low number of shared polymorphisms/haplogroups relative to exclusive polymorphisms is suggestive of NRY population structure, two additional analyses supported the hypothesis of strong population structure. AMOVA revealed that 43% of the total variance was partitioned among populations ($\Phi_{ST} = 0.43$; $P < 0.00001$) and population differentiation tests showed that all populations

TABLE 2

Amount and distribution of polymorphism within 26.5 kb by geographic region

Population	No. of pops.	n	S	π (%)	Θ (%)	Tajima's D	P	Fu and Li's D	P
All samples	35	92	46	0.0135	0.0341	−1.920	<0.05	−2.638	<0.05
Africans	7	28	20	0.0167	0.0194	−0.505	>0.10	−0.702	>0.10
Non-Africans	28	64	27	0.0085	0.0215	−1.916	<0.05	−2.294	>0.05
Asians	11	23	13	0.0066	0.0133	−1.774	>0.05	−1.339	>0.10
Europeans	7	20	10	0.0086	0.0107	−0.683	>0.10	−0.491	>0.10
Oceanians	3	8	6	0.0069	0.0083	−0.740	>0.10	−0.285	>0.10
Native Americans	7	13	6	0.0040	0.0073	−1.685	>0.05	−1.801	>0.10

were significantly differentiated ($P < 0.00001$; data not shown). The Φ_{ST} value for the 92 global samples grouped by continent was lower ($\Phi_{ST} = 0.32$), but still statistically significant ($P < 0.00001$).

Computer simulations: Coalescent simulation of samples drawn from a panmictic population experiencing exponential growth showed a dependence of TD on sample size (Figure 6). When the onset of growth occurred at $t_g = 10^3$ generations in the past, TD was less negative compared with $t_g = 5 \times 10^3$ generations ago. The rate of decrease of TD is initially high, and then TD begins to asymptote for sample sizes >100 chromosomes. When samples are divided into varying numbers of demes, the results systematically differ between the two models of population structure. Pooling samples generated by coalescent simulation of the island model of population structure did not produce the effect of an increasingly negative TD, either for 2 sampled chromosomes per deme (Figure 7, A and B) or for 20 chromosomes per deme (Figure 7, C and D). When migration in the island model is strong ($Nm = 1.0$), TD never significantly deviates from zero, even when the population experiences growth. Weak migration ($Nm = 10^{-3}$) in the island model produced simulated samples with positive TD values for each sampling protocol examined, thus countering the tendency of population growth to make TD negative. Although TD decreased as a function of the number of demes sampled when migration is weak (Figure 7, C and D), the values never became negative.

The population bifurcation model did produce TD values that became increasingly negative as samples were pooled from an increasing number of demes. When only two samples per deme were simulated, the mean TD was almost always negative and it became significantly negative when the number of demes pooled was >15 (Figure 8A). The only effect of population growth in this case was to decrease the values of TD slightly. When 20 chromosomes per deme were sampled, TD still became more negative as a greater number of demes were pooled (Figure 8B). When the total population size remained constant, TD never significantly differed

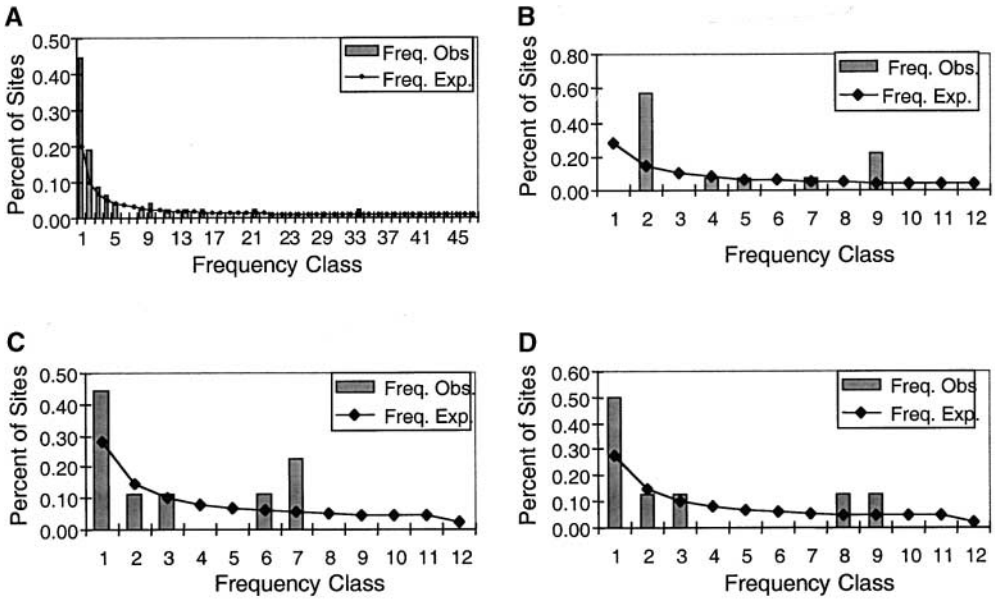


FIGURE 2.—Frequency spectra from a global sample of 92 chromosomes (A) and from three population-based samples: 25 Khoisan (B), 24 Papua New Guineans (C), and 24 Mongolians (D). Diamonds show the expected number of segregating sites in each frequency interval under a neutral, equilibrium model.

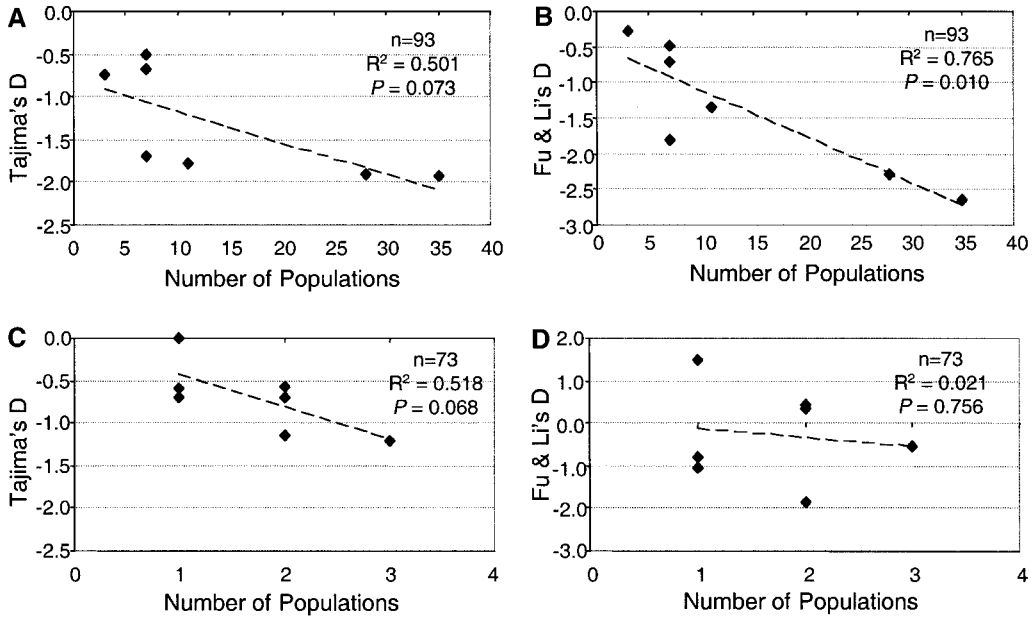


FIGURE 3.—Scatterplots of Tajima's *D* and Fu and Li's *D* *vs.* number of populations analyzed for the 26.5-kb noncoding region. (A) TD and (B) FD *vs.* number of populations in the global sample of 92 chromosomes, for combined non-Africans and for each continental region (*i.e.*, Africans, Asians, Europeans, Oceanians, and Native Americans). (C) TD and (D) FD *vs.* number of populations in the sample of 73 chromosomes from three populations ($n = 3$), for pairwise combinations of populations ($n = 2$) and for each population ($n = 1$; Khoisan, Papua New Guineans, and Mongolians).

from zero. However, when growth occurred, the values of TD from a sample of 20 chromosomes per deme were nearly identical to the values when 2 chromosomes per deme were sampled.

DISCUSSION

This study was motivated, in part, by the observation that patterns of NRY sequence variation differ according to sampling strategy. For example, early studies of Y chromosome polymorphism, based on sample sizes ranging from 5 (WHITFIELD *et al.* 1995) to 16 (HAMMER 1995), yielded TD values close to zero (Table 1). As technology to screen the NRY for polymorphisms became more efficient (UNDERHILL *et al.* 1997), longer lengths of DNA were screened in larger numbers of samples. SHEN *et al.* (2000) surveyed an average of 15 kb from each of four genes on the NRY in global samples of 53–72 males and found TD values that were consistently negative (TD < -1.5) and statistically significant ($P < 0.05$; Table 1; Figure 1). Here we screened 26.5 kb of noncoding NRY DNA for polymorphism in a global panel of 92 chromosomes and also found a statistically significant negative TD value ($P < 0.05$). Moreover, we demonstrate that TD and FD values are negatively correlated with sample size and the number of populations sampled, but were not correlated with the length of region examined.

Under the standard neutral model, the average value of TD or FD does not depend on sample size and has an expectation of approximately zero (TAJIMA 1989a; FU and LI 1993; SIMONSEN *et al.* 1995; FU 1996). Viola-

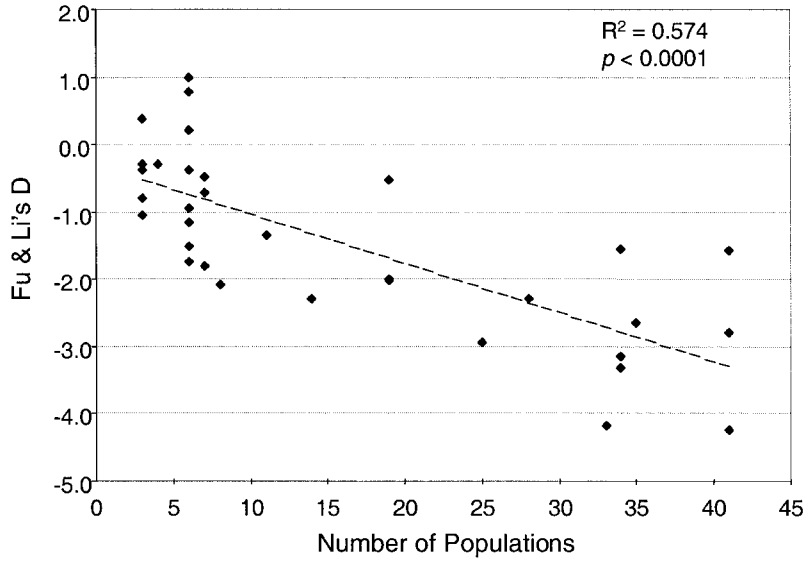


FIGURE 4.—Scatterplot of Fu and Li's *D* *vs.* number of populations analyzed for the combined data from the 26.5-kb noncoding region and four NRY genes (SHEN *et al.* 2000; see text). FD is shown for the entire global sample, for combined non-Africans, and for each continental region (*i.e.*, Africans, Asians, Europeans, Oceanians, and Native Americans).

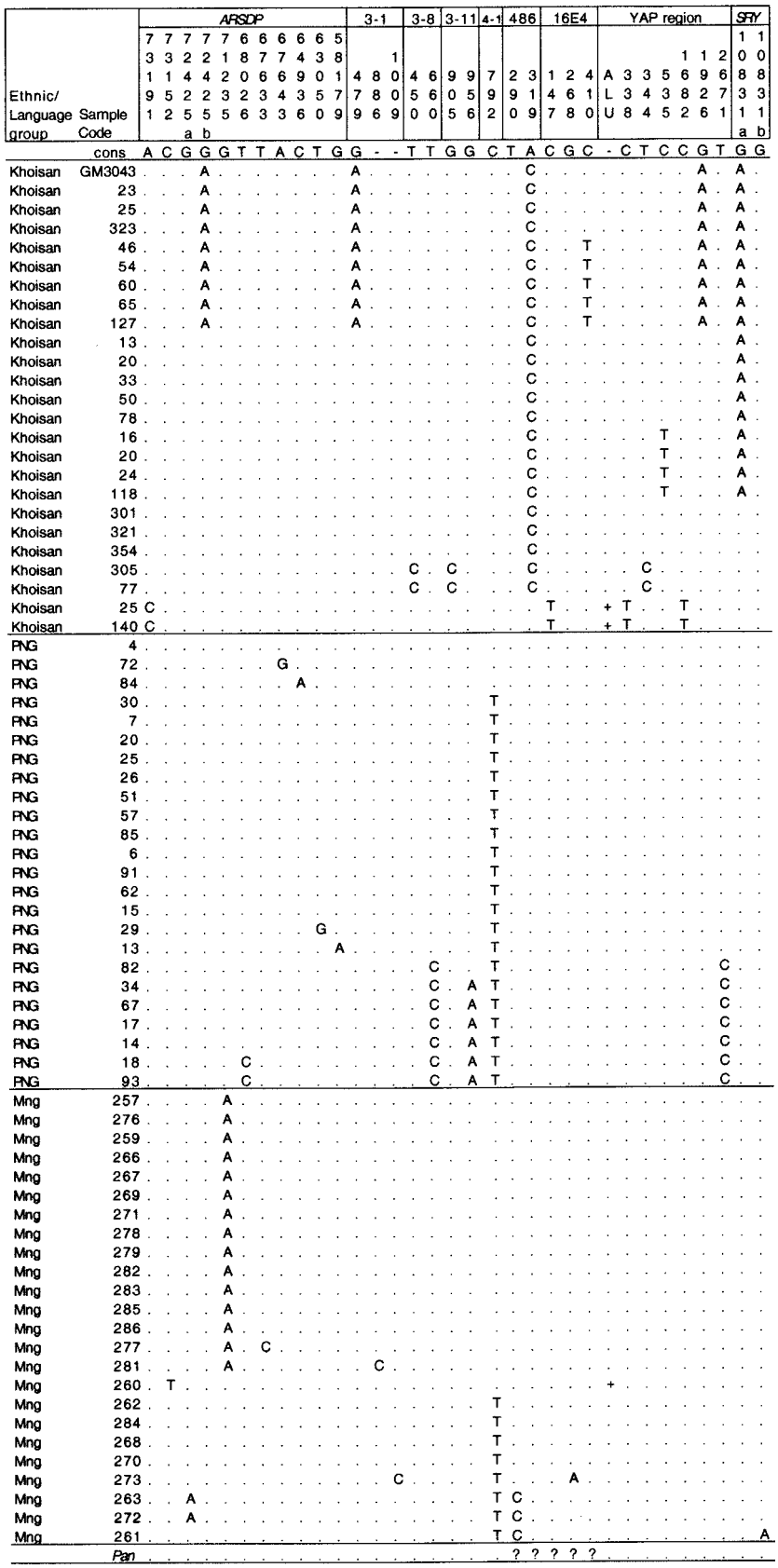


FIGURE 5.—Polymorphic sites in 26.5 kb of noncoding DNA on 73 Y chromosomes from three populations: the Khoisan, Papua New Guineans (PNG), and Mongolian Khalks (Mng). The numbering at the top of the figure is the same as in Figure 1.

TABLE 3

Amount and distribution of polymorphism within 26.5 kb in three populations

Population	*n*	No. of pops.	*S*	π (%)	Θ (%)	Tajima's *D*	*P*	Fu and Li's *D*	*P*
All samples	73	3	30	0.0144	0.0233	−1.208	>0.10	−0.528	>0.10
Khoisan	25	1	14	0.0140	0.0140	−0.007	>0.10	1.512	<0.02
PNG	24	1	9	0.0075	0.0091	−0.592	>0.10	−0.811	>0.10
Mongolians	24	1	8	0.0063	0.0081	−0.701	>0.10	−1.053	>0.10
Khoisan and PNG	49	2	23	0.0161	0.0195	−0.564	>0.10	0.417	>0.10
Khoisan and Mongolians	49	2	22	0.0146	0.0186	−0.695	>0.10	0.354	>0.10
PNG and Mongolians	48	2	16	0.0086	0.0135	−1.150	>0.10	−1.871	>0.05

tions of any one of the assumptions of the standard neutral model may underlie the observed negative TD and FD values. In the following sections we consider population structure and population growth as possible causative factors. Violations of the infinite sites mutation model are unlikely to be a major concern because parallel mutations are easily detected on the nonrecombining portion of the Y chromosome (Hammer *et al.* 1998; Underhill *et al.* 2001). The effects of directional selection are theoretically similar to those of population growth (Slatkin and Hudson 1991) and will be considered in a separate article.

Population structure: We posit that a violation of the assumption of random mating could give rise to the observed pattern on the Y chromosome, if a small number of individuals (*e.g.*, 1–3) are sampled from many locally differentiated populations. Recently, Ptak and Przeworski (2002) found that global sampling strategies lead to lower TD values than do population-based strategies and demonstrated that the principal factor influencing these lower TD values was the number of ethnicities (*i.e.*, populations) pooled in a sample. The NRY may be particularly prone to this type of sampling effect because many NRY polymorphisms are geographically localized (Jobling and Tyler-Smith 1995; Seielstad *et al.* 1998; Hammer *et al.* 2001). Thus, an increase in the number of populations surveyed will tend to cause an increase in the number of singleton or rare polymorphisms in the global sample. Indeed, we found very high levels of NRY population structure in this study, illustrated by the very low ratio of shared to exclusive polymorphisms (Figure 5) and by our finding of highly significant Φ_{ST} values. The results of analyzing subsets (*i.e.*, by continent) of global samples [using both our present data and those of Shen *et al.* (2000)] support the hypothesis that global sampling leads to biased estimates of the frequency of polymorphisms with singleton (or rare) variants.

To further explore the effects of sampling on the frequency distribution, we sequenced the same 26.5-kb noncoding region in a population-based sample of 24 or 25 chromosomes each from three human populations. We found more positive overall TD and FD values for the 73 sampled chromosomes from three populations than for the globally sampled set of 92 Y chromosomes (Table 3). The effect of accumulating singletons in the globally based sample of 92 Y chromosomes was shown by the stronger negative correlation using FD *vs.*

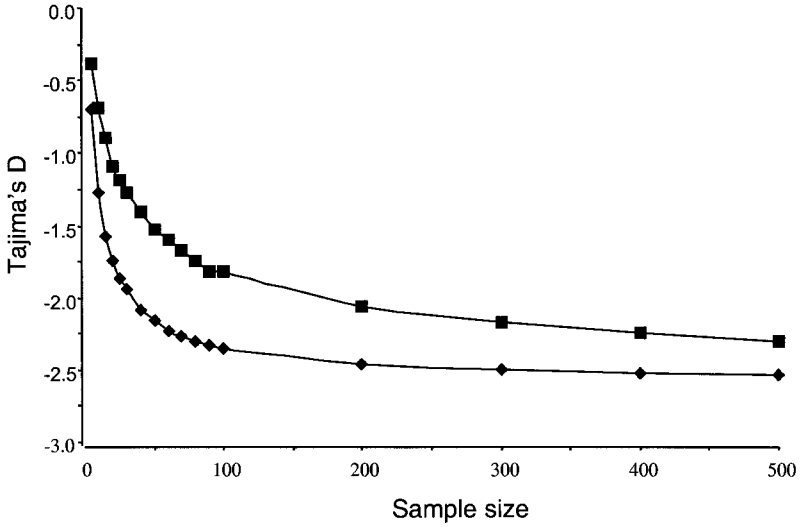


Figure 6.—The simulated Tajima's *D* as a function of the sample size taken from an exponentially growing panmictic population. Squares represent the values when population growth is assumed to have started 10^3 generations in the past. Diamonds represent the values of TD when growth began 5×10^3 generations in the past. Each value is the mean of 1000 replicate coalescent simulations.

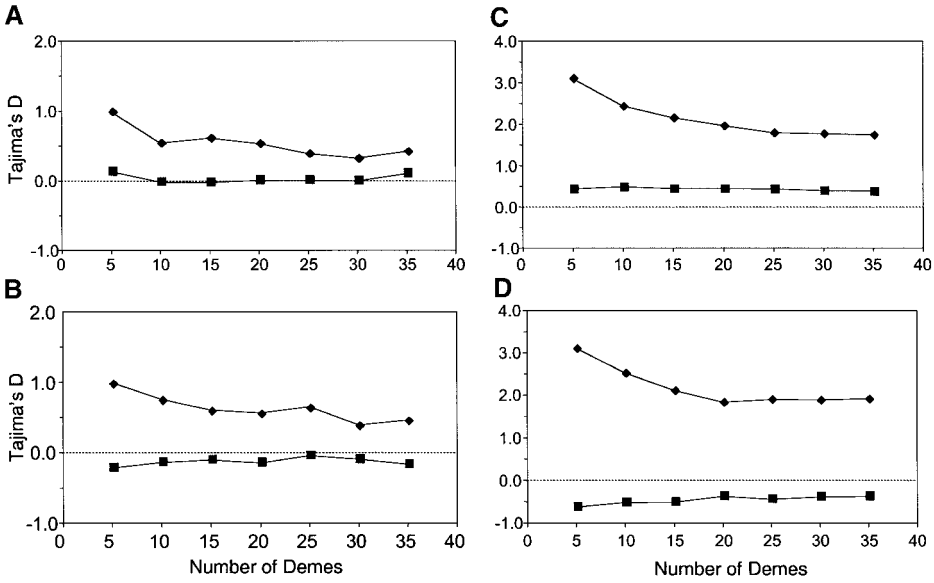


FIGURE 7.—The variation in simulated Tajima's *D* value as a function of the number of demes sampled under the island model of population structure. For each bout of simulations, two values of the migration parameter, *Nm*, were used (diamonds, $Nm = 0.001$; squares, $Nm = 1.0$). A and B show the simulated Tajima's *D* values when only 2 individuals per deme are taken. Values in A are for a constant size population, while those in B include exponential growth (see text). C and D show the Tajima's *D* values when 20 haploid individuals per deme are sampled. Values in C are for a constant size population and those in D include exponential growth.

TD as the number of subpopulations sampled increased. This difference is due to the stronger sensitivity of FD than of TD to singletons (FU and LI 1993). In contrast, the population-based sampling of 73 Y chromosomes produced a negative correlation with TD, but not with FD (Figure 3, C and D). This signifies that sampling only 1–3 individuals from many differentiated populations biases the frequency distribution toward polymorphisms with singleton variants. However, as one increases the number of samples per population and pools fewer numbers of differentiated populations, the bias toward singletons is diminished. Still, the skew toward rare polymorphisms may persist if there is a high ratio of exclusive to shared polymorphisms.

Computer simulations were also employed simply to establish that some models of population structure could lead to the observed relationship between TD and the number of populations sampled. We simulated samples under both the finite island model of population structure and a model of population bifurcation with no migration between demes. The island model of population structure, with weak migration, primarily produced samples with positive values of TD. TAKAHATA (1991) showed that, at the weak migration limit, most recent common ancestral lineages are reached quickly within demes, while the time required for these single ancestral lineages to coalesce between demes is determined by the low rate of migration. These long waiting times for interdemic coalescent events lead to a large number of fixed mutations between demes and therefore elevate the number of nucleotide differences in the total sample, as was shown initially by TAJIMA (1989b). When migration is weak, mutations will often be fixed within single demes and, as demes are pooled, TD is expected to decrease because the frequency of these exclusive mutations is decreased (Figure 7, C and D). However, at the weak migration limit, the long waiting time for interdeme coalescent events will also lead to a large number of shared mutations among demes. It is unclear whether TD will eventually become negative under the island model with a number of sampled demes similar to those typically sampled in empirical studies of humans. When migration is strong, there is little dependence of TD on the number of demes sampled or on the number of chromosomes sampled per deme (Figure 7).

The population bifurcation model of structure did produce samples with TD values that are compatible with our NRY data. Simulations of the global sampling strategy under this model of population structure yielded a negative correlation between TD and the number of pooled demes, whether or not growth was implemented. Analytical work examining the predictions of this type of population structure model is scant compared with the island model [although see TAKAHATA and NEI (1985) or WAKELEY and HEY (1997)] and no formal proof that this model systematically produces an excess of rare mutations is provided here apart from the simulation results. Yet, it is clear that a hierarchical isolation model appears to provide a better fit to the global NRY data than does an island model with either weak or strong migration.

Population growth: Undoubtedly, human populations have experienced both population growth and population structure at some time in the past. The question is to what extent either or both of these aspects of population history left a signature on patterns of variation. Under a growth model there is a dependence of TD and FD on sample size (Figure 6; PTAK and PRZEWORSKI 2002). To determine whether population growth rather than population structure could be driving the observed pattern of NRY variation, we reanalyzed

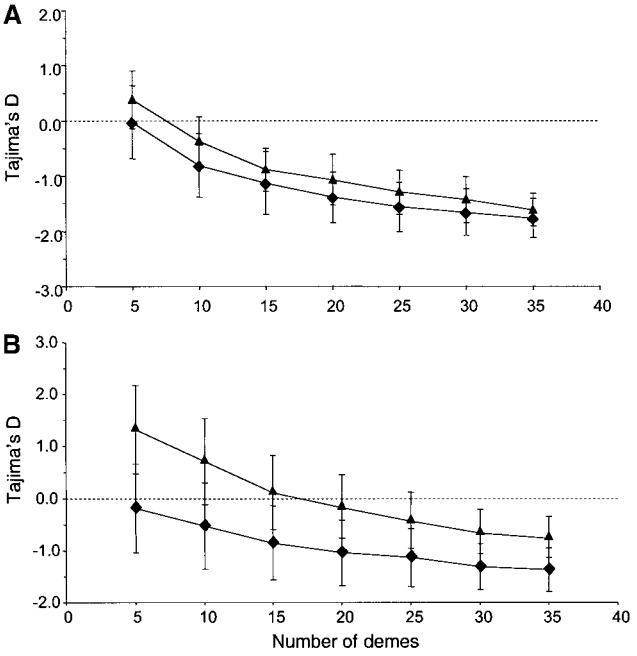


FIGURE 8.—The variation in simulated Tajima's *D* value as a function of the number of demes sampled under a population bifurcation model of structure. A shows the simulated Tajima's *D* values when two individuals from each deme are drawn. Triangles represent data for constant size populations, while the data denoted by diamonds include exponential population growth (see text). B shows the simulated Tajima's *D* values when 20 haploid individuals per deme are sampled. Triangles denote a constant size population, while the data represented by diamonds include exponential growth.

our population-based data by subsampling pools of demes of the same size as the individual population samples (*i.e.*, 8 individuals from each of three populations and 12 individuals from each of two populations). This strategy helps to control for sample size as a confounding variable (*i.e.*, under a growth model) leading to increasingly negative TD values in pooled samples. Randomly sampling 8 individuals from each of three populations (*i.e.*, the Khoisan, PNG, and Mongolians) and 12 individuals from each of two populations to produce random samples of $n = 24$ resulted in TD values similar to those for the pool of all three populations (*i.e.*, $n = 73$) and the average for paired populations (*i.e.*, $n = 48$), respectively (Table 4). Both of these TD values were more negative than the average TD value for the three individual populations (*i.e.*, $n = 24$ or 25; Table 4). Analyses of FD yielded similar patterns, albeit the random samples produced more negative values than those for any of the observed pooled data (Table 4). These results support the hypothesis that population structure is the chief factor underlying the negative trend in TD values, with little apparent effect of population growth.

As previously mentioned, simulation of the population bifurcation model shows that population growth is not necessary to create a negative correlation between TD and the number of pooled demes. Indeed, the addition of growth into the two models of population structure that we tested did not greatly influence the results of our simulations. In the island model, growth has only a negligible influence on the simulated values of TD, as can be seen by comparing Figure 7, A and C, with Figure 7, B and D, respectively. Likewise, in the population bifurcation model of structure, when only two chromosomes per deme are sampled, the influence of growth on TD is weak (Figure 8A). However, when 20 chromosomes per deme are sampled, the effect of growth is more pronounced, making TD more negative (Figure 8B). This suggests that if one wishes to distinguish the effects of population subdivision from population growth in a global sample, one must sample thoroughly within demes to obtain a robust estimate of the frequency distribution of mutations.

Implications for NRY studies: The observation of an excess of rare variants (*i.e.*, over those expected under a neutral, equilibrium model) in global NRY data sets has played a key role in supporting the hypothesis of a human Pleistocene population explosion (HARPENDING *et al.* 1998; HARPENDING and ROGERS 2000; SHEN *et al.* 2000; THOMSON *et al.* 2000). These results are particularly interesting in light of surveys of sequence variation that do not show a strong skew in the frequency spectrum toward rare alleles (WALL and PRZEWORSKI 2000; FRISSE *et al.* 2001; PTAK and PRZEWORSKI 2002; MARTH *et al.* 2003). Because of its smaller effective population size relative to the X chromosome and autosomes, one might expect that the NRY would be more influenced by *recent* population expansion(s) (FAY and WU 1999; PTAK and PRZEWORSKI 2002). The fact that the individual populations we surveyed do not show a statistically significant excess of rare NRY polymorphisms suggests that population subdivision (and not growth) may be responsible for the skew in the frequency spectrum often observed in global samples of NRY sequence variation.

It should be pointed out that the differences in frequency spectra patterns observed among the populations sampled here may be exacerbated by *local* population growth, decline, or selection. For example, the Khoisan are thought to have experienced a population contraction over the past several thousand years as a result of encroachment by expanding Bantu-speaking populations (EXCOFFIER *et al.* 1987; EXCOFFIER and SCHNEIDER 1999). Indeed, the frequency spectrum of the Khoisan is consistent with a decline in population size. It is also important to consider that different human populations experience changes in population size related to particular lifeways. For example, the Khoisan represent a small percentage of contemporary populations who are hunter-gatherers, while Mongolians and Papua New Guineans practice a mixture of foraging/herding and small-scale horticulture, respectively (CAVALLI-SFORZA *et al.* 1994). Thus, not all populations are expected to show the ef-

TABLE 4

TD and FD in single and pooled population samples from $n = 73$ data set

	n	TD ± SE	FD ± SE
Observed			
Average single population	24.3	−0.433 ± 0.076	−0.117 ± 0.162
Average paired populations	48.7	−0.803 ± 0.044	−0.367 ± 0.151
Three populations pooled	73.0	−1.208	−0.528
Random samples[a]			
3 × 8 (Khoisan/PNG/Mng)	24.0	−1.137 ± 0.079	−1.185 ± 0.162
2 × 12 (Khoisan and PNG)	24.0	−0.630 ± 0.076	−0.380 ± 0.151
2 × 12 (Khoisan and Mng)	24.0	−0.754 ± 0.071	−0.670 ± 0.151
2 × 12 (PNG and Mng)	24.0	−0.954 ± 0.069	−1.098 ± 0.119
2 × 12 (average)	24.0	−0.780 ± 0.033	−0.716 ± 0.074

PNG, Papua New Guineans; Mng, Mongolians.

[a] For each row 100 random samples were generated by choosing 8 individuals from each of three populations (or 12 from each of two populations) with replacement.

fects of population expansion, which may be more characteristic of populations that began practicing agriculture at the beginning of the Neolithic (Excoffier and Schneider 1999).

Furthermore, the scale of the observed pattern of NRY structure is not clear. Global surveys of NRY SNP variation show statistically significant structure both among continents and among populations within continents, as well as isolation by distance at some regional scales (Hammer *et al.* 2001). Therefore, additional sequence studies of thoroughly sampled populations from each continent are needed to infer the scale of subdivision on the NRY.

Finally, the results presented here suggest that estimates of the time of onset of population growth and the time to the most recent common ancestor (TMRCA) that are based on global sampling strategies and the assumption of panmixia should be considered with caution (Shen *et al.* 2000; Thomson *et al.* 2000). Models of exponential growth from a stationary panmictic population ignore the possible effects of population structure (Ptak and Przeworski 2002). This leaves two challenges for future models: (1) How do we correct for the sampling design to make reasonable estimates of growth rates and (2) how do we incorporate the effects of subdivision and population expansion into models to infer TMRCA and expansion times?

Conclusions: While the NRY is more susceptible to genetic drift and the effects of social processes (*e.g.*, patrilocality, polygyny, and/or kin-structured migration) that tend to increase the proportion of among-group variation, there is accumulating evidence of population structure affecting other regions of the genome. In the largest survey of sequence variation from a single panel of humans performed to date, Stephens *et al.* (2001) show that the majority of polymorphisms at 313 human autosomal and X-linked genes are restricted to individual populations, rather than shared among populations. Ptak and Przeworski (2002) have further demonstrated for autosomal and X-linked loci that the number of populations pooled in a sample has a significant negative correlation with TD. Thus, it appears that an excess of low-frequency polymorphisms may be present in pooled population data sets from all compartments of the genome. The confounding influence of population structure may persist even when only a single population is sampled, because rare alleles may still enter that population through migration (Ptak and Przeworski 2002). In this respect, it is intriguing that no studies that have deeply sampled single populations at autosomal loci (*e.g.*, Frisse *et al.* 2001; Koda *et al.* 2001), nor our own population samples from the NRY, have recovered an excess of rare alleles. Thus, while it appears that human population structure has left its signature on the genome in the frequency distribution of alleles recovered in global samples, neither the influence of growth nor migration among subdivided populations is evident when individual populations are surveyed more thoroughly. Interpreting these results in light of historical patterns of human subdivision, migration, and growth remains a challenging task.

We thank Ian J. Wilson for sharing computer code to help perform the Bayesian coalescent simulations of the population bifurcation model, Amit Indap for writing Perl scripts, and Himla Soodyall and Trefor Jenkins for providing Khoisan DNA samples. We also thank two anonymous reviewers for helpful suggestions. Publication of this article was made possible by grant GM-53566 from the National Institute of General Medical Sciences (to M.F.H.). Its contents are solely the responsibility of the authors and do not necessarily represent the official views of the NIH.

LITERATURE CITED

Allen, B. S., and H. Ostrer, 1994 Conservation of human Y chromosome sequences among male great apes: implications for the evolution of Y chromosomes. J. Mol. Evol. **39:** 13–21.

Alonso, S., and J. A. Armour, 2001 A highly variable segment of human subterminal 16p reveals a history of population growth

for modern humans outstide Africa. Proc. Natl. Acad. Sci. USA **98:** 864–869.

Batzer, M. A., M. Stoneking, M. Alegria-Hartman, H. Bazan, D. H. Kass *et al.*, 1994 African origin of human-specific polymorphic Alu insertions. Proc. Natl. Acad. Sci. USA **91:** 12288–12292.

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.

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

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

Excoffier, L., B. Pellegrini, A. Sanchez-Mazas, C. Simon and A. Langaney, 1987 Genetics and history of sub-Saharan Africa. Yearb. Phys. Anthropol. **30:** 151–194.

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

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

Fu, Y. 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 tests of neutrality of mutations. Genetics **133:** 693–709.

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.

Hammer, M. F., 1994 A recent insertion of an Alu element on the Y chromosome is a useful marker for human population studies. Mol. Biol. Evol. **11:** 749–761.

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

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.

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.

Hammer, M. F., T. M. Karafet, A. J. Redd, H. Jarjanazi, S. Santachiara-Benerecetti *et al.*, 2001 Hierarchical patterns of global human Y-chromosome diversity. Mol. Biol. Evol. **18:** 1189–1203.

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

Harpending, H. C., 1994 Signature of ancient population growth in a low-resolution mitochondrial DNA mismatch distribution. Hum. Biol. **66:** 591–600.

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.

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

Hudson, R., 1990 Gene genealogies and the coalescent process. Oxf. Surv. Evol. Biol. **7:** 1–44.

International SNP Map Working Group, 2001 A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. Nature **409:** 928–933.

Jobling, M. A., and C. Tyler-Smith, 1995 Fathers and sons: the Y chromosome and human evolution. Trends Genet. **11:** 449–456.

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

Koda, Y., H. Tachida, H. Pang, Y. H. Liu, M. Soejima *et al.*, 2001 Contrasting patterns of polymorphisms at the ABO-secretor gene (*FUT2*) and plasma α(1,3)fucosyltransferase gene (*FUT6*) in human populations. Genetics **158:** 747–756.

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

Martinez-Arias, R., F. Calafell, E. Mateu, D. Comas, A. Andres *et al.*, 2001 Sequence variability of a human pseudogene. Genome Res. **11:** 1071–1085.

Maruyama, T., 1970 Effective number of alleles in a subdivided population. Theor. Popul. Biol. **1:** 273–306.

Nachman, M. W., 2001 Single nucleotide polymorphisms and recombination rate in humans. Trends Genet. **17:** 481–485.

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:** 1855–1864.

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.

O'Donovan, M. C., P. J. Oefner, S. C. Roberts, J. Austin, B. Hoogendoorn *et al.*, 1998 Blind analysis of denaturing high-performance liquid chromatography as a tool for mutation detection. Genomics **52:** 44–49.

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

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

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

Raymond, M., and F. Rousset, 1995 An exact test for population differentiation. Evolution **49:** 1280–1283.

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

Romualdi, C., D. Balding, I. S. Nasidze, G. Risch, M. Robichaux *et al.*, 2002 Patterns of human diversity, within and among continents, inferred from biallelic DNA polymorphisms. Genome Res. **12:** 602–612.

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

Schneider, S., D. Roessli and L. Excoffier, 2000 *Arlequin: A Software for Population Genetic Analysis*. Genetics and Biometry Laboratory, University of Geneva, Geneva.

Seielstad, M. T., E. Minch and L. L. Cavalli-Sforza, 1998 Genetic evidence for a higher female migration rate in humans. Nat. Genet. **20:** 278–280.

Shen, P., F. Wang, P. A. Underhill, C. Franco, W. H. Yang *et al.*, 2000 Population genetic implications from sequence variation in four Y chromosome genes. Proc. Natl. Acad. Sci. USA **97:** 7354–7359.

Shen, P., M. Buchholz, R. Sung, A. Roxas, C. Franco *et al.*, 2002 Population genetic implications from DNA polymorphism in random human genomic sequences. Hum. Mutat. **20:** 209–217.

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

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

Stephens, J. C., J. A. Schneider, D. A. Tanguay, J. Choi, T. Acharya *et al.*, 2001 Haplotype variation and linkage disequilibrium in 313 human genes. Science **293:** 489–493.

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

Tajima, F., 1989b DNA polymorphism in a subdivided population:

the expected number of segregating sites in the two-subpopulation model. Genetics **123:** 229–240.

Takahata, N., 1991 Genealogy of neutral genes and spreading of selected mutations in a geographically structured population. Genetics **129:** 585–595.

Takahata, N., and M. Nei, 1985 Gene genealogy and variance of interpopulational nucleotide differences. Genetics **110:** 325–344.

Thomson, R., J. K. Pritchard, P. Shen, P. J. Oefner and M. W. Feldman, 2000 Recent common ancestry of human Y chromosomes: evidence from DNA sequence data. Proc. Natl. Acad. Sci. USA **97:** 7360–7365.

Thorstenson, Y. R., P. D. Shen, V. G. Tusher, T. L. Wayne, R. W. Davis *et al.*, 2001 Global analysis of ATM polymorphism reveals significant functional constraint. Am. J. Hum. Genet. **69:** 396–412.

Tishkoff, S. A., E. Dietzsch, W. Speed, A. J. Pakstis, J. R. Kidd *et al.*, 1996 Global patterns of linkage disequilibrium at the CD4 locus and modern human origins. Science **271:** 1380–1387.

Underhill, P. A., L. Jin, A. A. Lin, S. Q. Mehdi, T. Jenkins *et al.*, 1997 Detection of numerous Y chromosome biallelic polymorphisms by denaturing high-performance liquid chromatography. Genome Res. **7:** 996–1005.

Underhill, P. A., G. Passarino, A. A. Lin, P. Shen, M. Mirazon Lahr *et al.*, 2001 The phylogeography of Y chromosome binary haplotypes and the origins of modern human populations. Ann. Hum. Genet. **65:** 43–62.

Vigilant, L., M. Stoneking, H. Harpending, K. Hawkes and A. C. Wilson, 1991 African populations and the evolution of human mitochondrial DNA. Science **253:** 1503–1507.

Wakeley, J., and J. Hey, 1997 Estimating ancestral population parameters. Genetics **145:** 847–855.

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

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

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

Wilson, I. J., W. E. Weale and D. Balding, 2003 Inferences from DNA data: population histories, evolutionary processes and forensic match probabilities. J. R. Stat. Soc. Ser. A Stat. Soc. **166:** 155–201.

Y Chromosome Consortium, 2002 A nomenclature system for the tree of Y chromosomal binary haplogroups. Genome Res. **12:** 339–348.

Yu, N., Z. Zhao, Y. X. Fu, N. Sambuughin, M. Ramsay *et al.*, 2001 Global patterns of human DNA sequence variation in a 10-kb region on chromosome 1. Mol. Biol. Evol. **18:** 214–222.

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: M. Aguadé