Population Growth of Human Y Chromosomes: A Study of Y Chromosome Microsatellites

Jonathan K. Pritchard, Mark T. Seielstad,†[1] Anna Perez-Lezaun,‡[2] and Marcus W. Feldman**

*Department of Biological Sciences, Stanford University; †Department of Organismic and Evolutionary Biology, Harvard University; and ‡Lab. d'Antropologia, Facultat de Biologia, Universitat de Barcelona, Barcelona, Spain

We use variation at a set of eight human Y chromosome microsatellite loci to investigate the demographic history of the Y chromosome. Instead of assuming a population of constant size, as in most of the previous work on the Y chromosome, we consider a model which permits a period of recent population growth. We show that for most of the populations in our sample this model fits the data far better than a model with no growth. We estimate the demographic parameters of this model for each population and also the time to the most recent common ancestor. Since there is some uncertainty about the details of the microsatellite mutation process, we consider several plausible mutation schemes and estimate the variance in mutation size simultaneously with the demographic parameters of interest. Our finding of a recent common ancestor (probably in the last 120,000 years), coupled with a strong signal of demographic expansion in all populations, suggests either a recent human expansion from a small ancestral population, or natural selection acting on the Y chromosome.

Introduction

Like mtDNA, the Y chromosome is uniparentally inherited and nonrecombining along most of its length. These characteristics have motivated a number of recent studies which seek to infer the history of the Y chromosome, addressing questions similar to those examined for mtDNA (Vigilant et al. 1991). Of particular interest is the time of the most recent common ancestor (MRCA) of human Y chromosomes, which has implications for the origin and dispersal of modern humans. This problem has recently attracted considerable attention, including both experimental studies (Dorit, Akashi, and Gilbert 1995; Hammer 1995; Whitfield, Sulston, and Goodfellow 1995; Underhill et al. 1997; Hammer et al. 1998), and statistical analyses of existing data sets (Donnelly et al. 1996; Fu 1996; Fu and Li 1996, 1997; Weiss and von Haeseler 1996; Tavaré et al. 1997; Wilson and Balding 1998).

Estimates of the age of the MRCA are heavily dependent on both the assumed demographic model and estimates of its component parameters (Brookfield 1997; Tavaré et al. 1997). In a sense, this point is obvious: estimates of MRCA times are of interest largely because of the information that they carry about the size and structure of ancestral populations.

For these reasons, it is important to explore possible population models for human populations in order to understand the implications of departures from the simplest Wright-Fisher model. In addition, estimating demographic parameters (under a particular model) is a more direct way of inferring the details of human history than is estimating MRCA times.

In this paper, we focus in particular on the possible role of population expansion in generating the patterns of variation in human Y chromosomes. It is well known that the global human population has undergone a dramatic recent expansion (Cavalli-Sforza, Menozzi, and Piazza 1994, p. 105; Harpending et al. 1998). Thus, a constant-sized population model seems dubious.

Previous work on human mtDNA variation (e.g., Merriwether et al. 1991) has found consistent evidence of population growth for most populations (but see Weiss and von Haeseler 1998). There is also evidence for population growth in patterns of variation at autosomal microsatellite loci (e.g., Kimmel et al. 1998). In contrast, however, data from the β-globin locus are consistent with a constant-sized population model (Harding et al. 1997).

We analyzed microsatellite variation at a set of eight Y chromosome tri- and tetranucleotide repeats, surveyed in a worldwide sample of 445 human males (Perez et al. 1997; Seielstad, Minch, and Cavalli-Sforza 1998). These loci are highly variable and thus permit the analysis of fairly complicated evolutionary models. While most of the previous analyses of Y chromosome data have assumed a constant-population-size model, we consider a more general model which permits recent exponential population growth from an ancestral population of fixed size. This model has previously been studied by Weiss and von Haeseler (1998) in a study of mtDNA variation.

For a series of populations, haplotype information and estimates of single-locus variability are used to obtain posterior probabilities of the model parameters. Our analysis assumes the stepwise mutation model for microsatellite evolution (Goldstein et al. 1995; Slatkin 1995). Wilson and Balding (1998) recently implemented a Markov chain Monte Carlo (MCMC) algorithm for analyzing microsatellite data under the stepwise mutation model with constant population size. Analyzing a

[1] Present address: Program for Population Genetics, Harvard School of Public Health, Boston, Massachusetts.

[2] Present address: Facultat de Ciencies de la Salut i de la Vida, Department Biologia Evolutiva, Universitat Pompeu Fabra, Barcelona, Spain.

Key words: human evolution, coalescence times, most recent common ancestor, population growth, Y chromosome, demographic parameters.

Address for correspondence and reprints: Jonathan Pritchard, Department of Statistics, University of Oxford, 1 South Parks Road, Oxford, OX1-3TG, United Kingdom. E-mail: pritch@stats.ox.ac.uk.

Mol. Biol. Evol. 16(12):1791–1798. 1999

data set of (primarily) East Anglian (U.K.) chromosomes under a constant-sized population model, they reported that the 95% confidence interval of MRCA times extended from 15,000 to 130,000 years, with a mode at around 30,000 years, and that the effective population size was in the low thousands. Here, we consider a more general population and mutation model and a more geographically diverse data set.

Materials and Methods

Y chromosomes from 445 individuals representing 50 populations or ethnic groups were studied. The populations were classified into eight geographic clusters, listed below. These regional groups are also pooled into a worldwide sample (WORLD), and into African (AFR), and non-African (NonAFR) groups. The number of individuals representing each region or population and abbreviations for the regional names are given in parentheses.

Africans (AFR) (229)—East/Central Africa (EAFR) (113): Bench (8), Berta (8), CAR Pygmy (20), Dasenech (5), Dizi (4), Hamar (5), Konso (8), Majangir (10), Lissongo (4), Nyangatom (11), Ongota (9), Surma (11), Tsamako (5), Zaire Pygmy (5); Southern Africa (SAFR) (85): San (29), Sotho (17), Swazi (5), Tswana (12), Xhosa (7), Zulu (15); Western Africa (WAFR) (31): Bozo (4), Dogon (7), Peulh (6), Songhai (5), Tuareg (9).

Non-Africans (NonAFR) (216): Americas (AMER) (40): Colombia (5), Karitiana (9), Maya (7), Moskoke (3), Quechua (4), Surui (4), Ticuna (8); East Asia (EASIA) (46): Cambodia (16), China (4), Japan (10), Taiwan (15), Okinawa (1); Europe (EUR) (46): Basque (27), Catalan (14), Italian (2), German (3); Oceania (OCEAN) (24): Australia (6), Melanesia (6), New Guinea (12); West Asia (WASIA) (60): Baluchi (6), Brahui (6), Burushaski (24), Pathan (9), Sindhi (15).

Microsatellite Loci

All 445 individuals were typed at each of eight Y chromosome microsatellite markers: two trinucleotide repeats, DYS388 and DYS392, and six tetranucleotide repeats, DYS19, DYS389I, DYS389II, DYS390, DYS391, and DYS393 (Perez-Lezaun et al. 1997). Since the PCR fragment DYS389II contains the DYS389I fragment plus an additional microsatellite locus (de Knijff et al. 1997), we replaced the fragment size DYS389II by (DYS389II − DYS389I) in our analysis. None of these loci amplified in females.

Microsatellite Mutation Models

Pedigree studies of microsatellite mutation (e.g., Weber and Wong 1993) show that most mutations increase or decrease repeat scores by a single repeat unit; rarer mutations change the repeat score by two or more steps. Unless otherwise stated, we modeled the change in repeat score due to a single mutation using a symmetric geometric distribution with parameter p; we assumed a constant mutation rate μ, and we assumed that the probability of a mutation of size k at a single locus during one generation was $p(1 - p)^{|k|-1}\ \mu/2$ for $k \neq 0$, and $1 - \mu$ for $k = 0$. Except in model D, below, both μ and p were independent of the initial repeat score. Most previous theoretical work on microsatellites has been parametrized in terms of the variance of the mutation size distribution (assuming $k > 0$), σ^2. Since it is easy to compute the value of p corresponding to a given value of $\sigma^2 \geq 1$, henceforth our model will be specified in terms of σ^2.

A series of mutation models was used in order to investigate sensitivity to the model and parameters. These models were: (A) mutations drawn from a symmetric geometric as above, (B) pure one-step mutation (i.e., $p = 1$), (C) mutation rates variable among loci (with mutations drawn from a symmetric geometric as above), and (D) range constraints on maximum and minimum allowable repeat scores under the single-step mutation model. Some further details are given below. We did not consider the effect of asymmetric mutation rates.

In model C, the average mutation rate was set as some μ, and the mutation rate at the ith locus was set to $\mu V_i/\bar{V}$, where V_i and $\bar{V}$ were the observed variance at the ith locus and mean observed variance over all loci, respectively. In model D, we used pure stepwise mutation with reflecting boundaries (Feldman et al. 1997). The common ancestral allele at each locus was chosen to be j repeats, and the maximum and minimum repeat scores were $j + 3$ and $j - 3$, respectively. (For comparison, the average observed range was slightly larger, at 6.5 repeats. We chose an allowable range of 6.0 repeats so that the effect of range constraints, if any, would be less extreme than under this model.)

Population Model

We assume that human population histories can be represented by a model of exponential growth from an ancestral population of fixed size. That is, we assume that there was a random-mating ancestral population of constant size containing an (effective) number N_A of Y chromosomes. At time t_0/g generations before the present, the population began exponential growth at rate $r \geq 0$ per generation. (Here, t_0 is the time in years at which exponential growth started, and g is the average generation time.) The modern effective population size is then $N_A e^{(rt_0/g)}$. This model is closely related to that assumed by Weiss and von Haeseler (1998).

Prior Values

The two largest data sets of microsatellite mutations (Weber and Wong 1993; Dib et al. 1996) suggest that dinucleotide mutation rates are roughly 6×10^{-4} per locus per generation. Furthermore, Chakraborty et al. (1997) estimated that autosomal tetranucleotide rates are roughly half of dinucleotide rates. In contrast, Heyer et al. (1997), who studied Y chromosome microsatellite mutation rates directly from pedigrees, estimated a rather higher rate, 2×10^{-3}, albeit on the basis of just a few observed mutations. In order to model the uncertainty in the value of μ, we used a Gamma prior with parameters $(10, 12{,}500^{-1})$. This distribution has a mean at 8×10^{-4} and permits both of the above estimates.

We modeled the variance in mutation size, σ^2, as having a prior distribution of 1.0 + Expo(1.0), where Expo(λ) is an exponentially distributed random variable with mean λ. (The prior mean value for σ^2 of 2.0 corresponds to a value of $p = 0.78$ in the mutation model). An estimate of σ^2 near 2.0 was suggested by the data of Dib et al. (1996).

Since the major goal of this study is estimation of the population parameters, and since the prior information on these is fairly vague, we adopted rather diffuse priors for these. Typical values of N_e estimated using the constant population size model are on the order of 10,000. Therefore, we used a rather vague prior for N_A, namely a LogNormal with parameters 8.5 and 2. The prior on r was Expo(0.005) per generation. The prior on t_0 was Expo(20,000) years ago, corresponding to a mean time for the start of population growth of 20,000 years ago. This exponential growth model with mutation model A will be referred to as the standard model. We assume that the generation time, g, is 20 years.

Because of our interest in whether the data supported the exponential-growth model, we also present results from an additional set of simulations, in which we placed half the prior weight on a model of constant population size, and half on the standard model described above. More specifically, we changed the prior on t_0 so that with probability 0.5, $t_0 = 0$ (this is, in effect, a model of constant population size), and with probability 0.5, $t_0 \sim$ Expo(20,000), as before. The other parameters were as above.

Estimation Procedure

Tavaré et al. (1997) describe a rejection algorithm for estimating common ancestor times for DNA sequence data when there is prior information about population demography. Their algorithm is designed for the infinite-sites mutation model, and it replaces the full data set by a summary statistic, the number of segregating sites. The method used here is derived from their approach but is less computationally efficient due to the more complicated microsatellite mutation process.

The amount and distribution of variation in a sample of m chromosomes were summarized by three statistics: $\bar{V}$, the mean (across loci) of the variance in repeat numbers, $\bar{H}$, the mean effective heterozygosity (i.e., the probability of two randomly drawn chromosomes differing at a particular locus, averaged across loci), and n, the number of distinct haplotypes. (The term ''haplotype'' refers to the vector containing the repeat scores at each of the eight loci on a particular chromosome. Two haplotypes are considered to be different if they contain different alleles at any one of the eight loci.)

These statistics were chosen because they exhibit different behaviors under different population models. The variance and heterozygosity can both be used to estimate $N\mu$ (Pritchard and Feldman 1996), and, used together, they contain some information about population size changes (Kimmel et al. 1998). Furthermore, since the Y chromosome is nonrecombining, it is sensible to include haplotype information in addition. We used the number of haplotypes in the sample because simulations have shown this to be strongly affected by the population history (data not shown). In fact, the number of haplotypes depends in large part on the lengths of the terminal and subterminal branches in the genealogy, because mutations must accumulate on those branches in order to create rare haplotypes. Thus, for a given amount of variation per locus, there will tend to be more distinct haplotypes in a growing population, which has longer terminal branches, than in a population of constant size.

For a data set of m chromosomes, with summary statistics $\bar{V}$, $\bar{H}$, and n, we estimated posterior distributions for the parameters of interest using the following rejection algorithm:

1. Simulate μ, σ^2, N_A, r, and t_0 independently from the prior distributions given above.
2. Simulate a sample of m chromosomes typed at eight microsatellite loci each using the simulated values of μ, σ^2, N_A, r, and t_0. The trees were generated using standard coalescent algorithms (Hudson [1990] and Slatkin and Hudson [1991] for the exponential growth phase). Microsatellite mutations were generated from the distributions described above in *Microsatellite Mutation Models.*
3. Compute $\bar{V}^*$, $\bar{H}^*$, and n^*, the variance, heterozygosity, and number of haplotypes in the simulated sample, respectively.
4. If all of $|\bar{V} - \bar{V}^*|/\bar{V}$, $|\bar{H} - \bar{H}^*|/\bar{H}$, and $|n - n^*|/n$ are less than a small number δ, then record μ, σ^2, N_A, r, t_0, and the simulated time of the MRCA.
5. Return to 1.

This procedure gives a sample from the posterior of the parameters, conditional on ($\bar{V}$, $\bar{H}$, n) being within δ of the observed values. As $\delta \to 0$, this corresponds to conditioning on the observed values; however, the algorithm becomes very inefficient.

After experimenting with different values of δ, the results presented use values of 0.1 or less. Smaller values of δ typically did not reduce the sizes of the credible intervals appreciably. All posterior results are based on at least 10^6 repetitions, and the acceptance rates were generally on the order of 10^{-3} for the models with exponential growth. For some populations, the acceptance rate was less than 10^{-6} under the constant-sized population model.

Results

In the overall sample, there were between 6 and 11 alleles at each locus. The mean heterozygosity (per locus) was 0.636, and the mean variance across loci in the number of repeat units was 1.149. The data used in the calculations are summarized in table 1.

Previous studies of autosomal microsatellites have found that African populations are typically more variable (in terms of heterozygosity or variance in repeat scores) than non-African populations (e.g., Bowcock et al. 1994; Deka et al. 1995; Jorde et al. 1995; but see also Harding et al. [1997], who observed a higher frequency of pairwise differences in Asia than in Africa).

Table 1
Levels of Genetic Diversity, by Geographic Region

	m	n	$\bar{V}$	$\bar{H}$
WORLD	445	316	1.149	0.6358
AFR	229	151	1.175	0.5948
NonAFR	216	169	0.900	0.6296
EAFR	113	80	0.965	0.5856
SAFR	85	60	1.358	0.5657
WAFR.	31	19	1.055	0.5930
AMER.	40	30	0.502	0.5300
EASIA	46	36	0.770	0.5822
EUR	46	35	0.646	0.4920
OCEAN.	24	21	0.939	0.6255
WASIA	60	50	0.916	0.6244

NOTE.—m = sample size; n = number of distinct haplotypes; $\bar{V}$ = mean (across loci) variance in repeat score; $\bar{H}$ = mean (effective) heterozygosity. See text for population abbreviations.

Our data do not show a clear excess of variability in African populations. The observed variance in repeat scores is higher within Africa than elsewhere, but the observed heterozygosity is lower.

Among the 445 chromosomes in the overall sample, there were 316 distinct haplotypes. Since these Y chromosome loci are nonrecombining, every novel haplotype must result from at least one mutation. This implies a large number of mutation events in the ancestral genealogy of the sample (at least 315), and since the average minimum number of mutations per locus far exceeds the number of alleles per locus, it is clear that particular alleles must have arisen independently many times.

Despite the inferred pattern of parallel mutation, there is significant linkage disequilibrium among many of the loci in the sample. Applying Fisher's exact test to the allele combinations at pairs of loci, we found that 16 of the 28 pairs were in significant linkage disequilibrium at the 5% level, using the (conservative) Bonferroni criterion to correct for multiple comparisons.

The presence of allelic associations can also be inferred from the number of distinct haplotypes. We used a series of random permutations to form new sets of haplotypes in which the allele frequencies were the same as in the original data, but with random assignment of alleles to chromosomes. The mean number of distinct haplotypes in the random sets was 426.5 ± 3.2, which is highly significantly different from the 315 observed. This randomization procedure mimics the distribution that would be expected from a perfect star phylogeny, for which there would be no linkage disequilibrium (since no mutations are shared by descent among lineages). We can therefore reject such an extreme model.

Summary of Results for World Sample

Table 2 summarizes the posterior results for the world sample under a variety of models. We used rather vague priors for all of the parameters (except μ); these are listed as "Pre-Data." In all cases (except that of μ), the posterior distributions are quite different from the prior, with much tighter bounds. This indicates that the data contain a lot of information about the parameters of interest.

As discussed below, we found strong support for the exponential growth model as compared with the model of constant population size. This can also be seen from the posterior estimates of the growth rate parameter r, which are of the order of 0.008 per generation under all models, and the fact that the lower bounds on the posterior interval for r are larger than they are in the prior.

Table 2
Estimates of Parameter Values for the World-wide Sample Under Various Models

	$\hat{N}_A$	$\hat{N}_A$ (range)	$\hat{r}$	r (range)
Standard	1,000	50–3,500	0.0080	0.0026–0.0226
$\sigma^2 \equiv 1$	1,500	100–4,900	0.0075	0.0022–0.0209
Variable μ	1,100	70–3,800	0.0076	0.0023–0.0204
Range constr.	2,000	200–6,500	0.0080	0.0021–0.0221
Pre-Data	36,000	100–231,000	0.0050	0.0001–0.0187
	$\hat{t}_0$	t_0 (range)	$\hat{T}$	T (range)
Standard	18,000	7,000–41,000	46,000	16,000–126,000
$\sigma^2 \equiv 1$	18,000	6,000–43,000	65,000	24,000–164,000
Variable μ	19,000	7,000–44,000	47,000	18,000–110,000
Range constr.	17,000	6,000–43,000	91,000	25,000–318,000
Pre-Data	20,000	600–75,000	1,240,000	8,000–9,251,000
	$\hat{\sigma}^2$	σ^2 (range)	$\hat{\mu}$	μ (range)
Standard	1.47	1.02–2.39	0.0007	0.0003–0.0012
$\sigma^2 \equiv 1$	1.00	1.00–1.00	0.0007	0.0004–0.0012
Variable μ	1.35	1.02–2.04	0.0007	0.0004–0.0012
Range constr.	1.00	1.00–1.00	0.0007	0.0004–0.0012
Pre-Data	2.00	1.03–4.85	0.0008	0.0004–0.0014

NOTE.—The estimates given are the means of the posterior distributions. "Range" refers to 95% probability intervals. "Standard" refers to the exponential growth model described in the *Population Model* and *Prior Values* sections. "Pre-Data" gives the corresponding priors. The other models shown (see B, C, and D, *Microsatellite Mutation Models* section) are "$\sigma^2 \equiv 1$" (pure stepwise mutation), "Variable μ" (relative mutation rates varied across loci), and "Range constr." (range constraints). T denotes the MRCA time of the samples. Both T and t_0 are given in years, assuming a generation time of 20 years; r is the growth rate per generation.

Table 3
Estimates of Parameter Values for the Various Subgroups of the Sample

	$\hat{N}_A$	$\hat{N}_A$ (range)	$\hat{r}$	r (range)
WORLD	1,000	50–3,500	0.0080	0.0026–0.023
AFR	900	50–3,200	0.0077	0.0024–0.022
NonAFR	800	60–2,700	0.0085	0.0027–0.022
EAFR	900	60–3,000	0.0066	0.0018–0.020
SAFR	1,000	80–3,400	0.0069	0.0017–0.019
WAFR	1,200	200–3,200	0.0028	0.0001–0.013
AMER	700	40–2,500	0.0060	0.0009–0.016
EASIA	900	70–3,200	0.0056	0.0010–0.017
EUR	600	40–2,200	0.0075	0.0016–0.020
OCEAN	1,200	80–4,200	0.0051	0.0004–0.016
WASIA	1,000	60–3,300	0.0062	0.0014–0.018
Pre-Data	36,000	100–231,000	0.0050	0.0001–0.019
	$\hat{t}_0$	t_0 (range)	$\hat{T}$	T (range)
WORLD	18,000	7,000–41,000	46,000	16,000–126,000
AFR	15,000	5,000–37,000	44,000	14,000–128,000
NonAFR	17,000	7,000–38,000	38,000	15,000–94,000
EAFR	15,000	5,000–38,000	40,000	14,000–118,000
SAFR	13,000	3,000–32,000	45,000	13,000–137,000
WAFR	12,000	200–49,000	44,000	14,000–125,000
AMER	14,000	3,000–37,000	27,000	10,000–66,000
EASIA	15,000	4,000–40,000	36,000	13,000–92,000
EUR	13,000	4,000–32,000	27,000	10,000–74,000
OCEAN	16,000	1,000–45,000	42,000	15,000–107,000
WASIA	17,000	5,000–43,000	40,000	14,000–107,000
Pre-Data	20,000	600–76,000	1,240,000	8,000–9,251,000

NOTE.—Notation as in table 2. The priors used are as given in the *Prior Values* section; the Pre-Data values reported for T are for the entire sample of 445 chromosomes. See text for population abbreviations.

The estimates of ancestral population size N_A (less than about 5,000) are close to the sorts of values often estimated under the assumption of constant population size (e.g., Wilson and Balding 1998). However, such estimates of ancestral population size seem more realistic here than in the constant-population-size model, given that our results also indicate that the last 6,000 years or more (t_0) have been accompanied by substantial population growth.

As usual in this type of analysis, there is considerable posterior uncertainty with regard to the time of the MRCA (T). Most of the support here is for a relatively recent ancestor (in the last 120,000 years or so), although the model with range constraints permits a rather older MRCA. In converting these estimates from generations to years, we did not consider uncertainty in the generation time, but instead assumed a generation time of 20 years. The effect of a different generation time would be a proportional rescaling of the time estimates.

Table 2 also contains information about the variance in the size of microsatellite mutations, σ^2. A common problem in using microsatellites for dating is that while not very much is known about σ^2, the estimates depend heavily on the assumed value (Feldman, Kumm, and Pritchard 1999). Our approach here was to set a relatively vague prior on σ^2. The combination of summary statistics (especially variance and heterozygosity) then contains some information about σ^2. We found that the single-step mutation model produces a good fit to the data, while a mutation process with large variance can be rejected. Our data suggest that for these loci the value of σ^2 is between 1.0 and 2.4, with a posterior mean of about 1.4.

Sensitivity to the Mutation Model

Since there is considerable uncertainty about the details of the microsatellite mutation process, we considered a series of four alternative models (see *Microsatellite Mutation Models*). The goal was to characterize the degree to which our estimates depended on the assumed mutation process. We found that our estimates of r and t_0 were quite consistent across mutation models.

In contrast, estimates of the time of the MRCA and (to a lesser extent) ancestral population size were larger under the one-step mutation model and the range constraint model than under the other models. This is hardly surprising—by preventing large mutations, these models lower the expected values of each of the summary statistics (at any given set of parameter values). In principle, if μ were high enough, range constraints could obscure the signal of evolutionary history altogether; this does not appear to be the case for these data, given that (1) there is strong linkage disequilibrium among the loci, and (2) extreme values of the MRCA time and N_A permitted under the prior are excluded by the posterior for the range constraints model.

Summary of Results for Regional Populations

Table 3 summarizes the results obtained for the regional populations. While our results permit some general comments about parameter values in our model, the posterior intervals are too broad to draw firm conclu-

sions about the relative values of the parameters in the various populations.

All of the populations (except those of west Africa and Oceania, as discussed below), support a model of fairly sustained population growth over a period of some thousands of years, starting from a small ancestral population size, on the order of a few thousand Y chromosomes. For all of the populations, as well as the overall world sample, the time to the MRCA is estimated to be on the order of 40,000 years ago, but with a large range of uncertainty, up to about 120,000 years. As noted above, this range would increase if there were tight constraints on the possible repeat scores at the microsatellite loci.

The posterior estimates of the time at which population growth started are not highly informative, but they are consistent with an expansion associated with the start of agriculture some 10,000 years ago.

Comparison of Constant-Size and Exponential-Growth Models

In order to assess the degree of support for the exponential growth model compared with a model of constant population size, we considered a model in which 50% of the prior probability was placed on each model. Then, for a given δ, we estimated the acceptance rates for each model. (Further details are given at the end of the *Prior Values* section.)

For most of the populations, virtually all of the posterior support is on the exponential-growth model. We were able to reject a model of constant population size in all populations except those of west Africa and Oceania. The constant-size model has posterior probability $< 5\%$ for east Asia and America, and $< 1\%$ for the worldwide, African and non-African groupings, as well as for Europe, east Africa, southern Africa, and west Asia.

In separate simulations, we found that the model of constant population size typically produced a very poor fit to our data. Realistic values of $N\mu$ (based on the observed variance and heterozygosity) produced far fewer distinct haplotypes in a sample of size m than observed in most of our data. The better fit to the model of population growth is due to the fact that with population growth, the terminal branches of the genealogy tend to be longer than they are under the constant-population-size model, producing more distinct haplotypes for a given level of pairwise diversity.

As noted above, two populations (those of west Africa and Oceania) did not provide strong support for the population growth model. The posterior probabilities of the constant population size model were 0.63 and 0.16 for west Africa and Oceania, respectively. Also, for west Africa, the estimate of r under the exponential-growth model was considerably lower than those for the other populations. This strong support for the constant-population-size model in west Africa, combined with the low estimate of r, suggests that in this population, growth may have been weak or absent. This is in strong contrast to most of the other populations considered.

Discussion

The early Y chromosome data sets contained few haplotypes (e.g., Dorit, Akashi, and Gilbert 1995; Hammer 1995; Whitfield, Sulston, and Goodfellow 1995), restricting most analyses to a simple demographic model: constant population size. More recent data sets (e.g., Underhill et al. [1997] and microsatellite data sets such as those considered here) contain larger numbers of ancestral mutations and thus present the opportunity to investigate more complex models.

In this study, we considered a model of human demography which allows for population growth starting at some time t_0 in the past and continuing to the present. Although we followed most of the previous Y chromosome studies in estimating the time to the MRCA, we also estimated a series of demographic parameters under the assumed model. Estimates of MRCA times permit easy comparison among studies which use different methods or different loci. However, the MRCA time is only indirectly connected to population history and is notoriously difficult to estimate with high precision. For these reasons, more can be learned about the history of populations from estimates of demographic parameters.

The major finding of this study is that a model of constant population size produces a very poor fit to the data for most of the populations considered. The model of exponential growth performs far better, and for the worldwide sample, we obtained a point estimate for the population growth rate of 0.008 per generation, starting 18,000 years ago.

The consistent pattern of population growth indicates that genetic distance measures and statistical tests that assume constant population size are inappropriate for the human Y chromosome. It will be important to evaluate which distance measures and tree-building methods are robust in the presence of population growth. In addition, the neutral expectation that Y chromosomal loci should have ¼ as much variation as autosomal loci is not valid if there has been a consistent pattern of population growth. In that case, the levels of variation should be more similar (Slatkin and Hudson 1991).

It is interesting to compare our estimates of MRCA times (table 2) with those from several previous studies. We found evidence for a relatively recent common ancestor of the worldwide sample, with point estimates ranging between about 45,000 and 90,000 years ago, depending on the mutation model, but with considerable uncertainty around those estimates.

Underhill et al. (1997) made estimates for two sets of DNA sequence data. Their estimates were 162,000 years (range 69,000–316,000 years) and 186,000 years (range 77,000–372,000 years) using the method of Fu and Li (1997). Hammer et al. (1998) used genotype data from a series of nine biallelic sites, and estimated the time to the MRCA at 147,000 years (range 68,000–258,000 years) using a version of the method developed by Griffiths and Tavaré (1994). Their analysis was complicated by the fact that the segregating sites had been

identified previously, making it difficult to estimate the classical mutation parameter θ.

Both of these studies obtained estimates that are higher than those obtained here, except under our model D (tight range constraints), although our credible intervals overlap considerably with theirs. Both Underhill et al. (1997) and Hammer et al. (1998) assumed a constant population size of 5,000 individuals, considerably larger than our estimates for N_A (mean 1,000, range 50–3,500) and this difference may account in large part for the higher estimates that they obtained.

Wilson and Balding (1998) used a full Bayesian analysis to estimate an MRCA time of 30,000 years (range 15,000–130,000 years) using a set of microsatellite data of limited geographic origin. Their estimates are comparable with ours, although their analysis differed in that it assumed a constant population size, single-step mutation (see our table 2, model B), and a slightly higher mutation rate.

As noted above, our data do not show a clear excess of variation in Africa relative to the rest of the world, in contrast with other studies of microsatellite and sequence variation. All the populations (except possibly that of west Africa) show very similar patterns of population growth from a small ancestral population at approximately the same time. That is, our data do not support the idea that the effective population size of Y chromosomes is much larger in Africa than elsewhere, as has been found for other parts of the genome (e.g., Jorde et al. 1995; Stoneking et al. 1997).

A serious issue in population genetic studies using microsatellite variation is that the details of the microsatellite mutation process are only poorly understood. Most previous studies have assumed a particular version of the stepwise mutation process (usually setting $\sigma^2 \equiv 1$ and assuming no range constraints). Our approach here was to consider a number of possible mutation mechanisms and to investigate the extent to which the estimates depended on these assumptions. Furthermore, we chose to treat σ^2 as unknown and perform inference on it simultaneously with the demographic parameters of interest. This sort of approach should make our results more robust than those of studies in which the mutation process and σ^2 are assumed to be known.

Nevertheless, our analysis has several modeling limitations. First, while the demographic model considered here is considerably more general than a model of constant population size, we still ignore a number of features of realistic populations. In particular, we ignore population structure.

Second, we assume selective neutrality of the Y chromosome. It is difficult, a priori, to know whether the signal of population growth that we have seen here, coupled with the apparently recent MRCA time, is the result of neutral demographic processes or of natural selection. It has been found in other species, particularly in *Drosophila,* that regions of little or no recombination—like the human Y chromosome—frequently have very little genetic variation as a result of selection (e.g., Berry, Ajioka, and Kreitman 1991; Begun and Aquadro 1992). Recent results of Nachman et al. (1998) suggest a similar trend in humans. It would be of considerable biological interest if natural selection were shown to have been an important force on the human Y chromosome, but the value of the Y chromosome as a tool for interpreting human history would then be reduced. Of course, these comments apply to any loci, and ultimately a clear picture of human history will emerge only by combining information from many loci.

Acknowledgments

This work was supported by NIH grants GM28428 and GM28016. J.K.P. was supported by a Howard Hughes predoctoral fellowship. We thank M. Stephens, F. Stefanini, and the reviewers for helpful comments. The data reported in this paper are available from www.stats.ox.ac.uk/~pritch/home.html.

LITERATURE CITED

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

Berry, A. J., J. W. Ajioka, and M. Kreitman. 1991. Lack of polymorphism on the Drosophila fourth chromosome resulting from selection. Genetics **129**:1111–1117.

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

Brookfield, J. F. Y. 1997. Importance of ancestral DNA ages. Nature **388**:134.

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

Chakraborty, R., M. Kimmel, D. N. Stivers, L. J. Davison, and R. Deka. 1997. Relative mutation rates at dinucleotide, trinucleotide and tetranucleotide microsatellite. Proc. Natl. Acad. Sci. USA **94**:1041–1046.

De Knijff, P., M. Kayser, A. Caglià, D. Corach, N. Fretwell, B. Herzog, M. Hidding, K. Honda, M. Jobling, and M. Krawczak. 1997. Chromosome Y microsatellites: population genetic and evolutionary aspects. Int. J. Legal Medicine **110**:134–140.

Deka, R., L. Jin, M. D. Shriver, L. M. Yu, S. DeCroo, J. Hundrieser, C. H. Bunker, R. E. Ferrell, and R. Chakraborty. 1995. Population genetics of dinucleotide $(dC–dA)_n$-$(dG–dT)_n$ polymorphisms in world populations. Am. J. Hum. Genet. **56**:461–474.

Dib, C., S. Faure, C. Fizames, D. Samson, N. Drouot, and E. A. Vignal. 1996. A comprehensive genetic map of the human genome based on 5,264 microsatellites. Nature **380**: 152–154. [Extended reprint: A1–A138.]

Donnelly, P., S. Tavaré, D. J. Balding, and R. C. Griffiths. 1996. Estimating the age of the common ancestor of men from the ZFY intron. Science **272**:1357–1359.

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

Feldman, M. W., A. Bergman, D. D. Pollock, and D. B. Goldstein. 1997. Microsatellite genetic distances with range constraints: analytic description and problems of estimation. Genetics **145**: 207–216.

Feldman, M. W., J. Kumm, and J. K. Pritchard. 1999. Mutation and migration in models of microsatellite evolution. Pp. 88–115 *in* D. Goldstein and C. Schlotterer, eds.

Microsatellites: evolution and applications. Oxford University Press, Oxford, England.

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

FU, Y.-X., and W.-H. LI. 1996. Investigating the age of the common ancestor of men from the ZFY intron. Science **272**:1356–1357.

———. 1997. Estimating the age of the common ancestor of a sample of DNA sequences. Mol. Biol. Evol. **14**:195–199.

GOLDSTEIN, D. B., A. R. LINARES, L. L. CAVALLI-SFORZA, and M. W. FELDMAN. 1995. An evaluation of genetic distances for use with microsatellite loci. Genetics **139**:463–471.

GRIFFITHS, R. C., and S. TAVARÉ. 1994. Ancestral inference in population genetics. Stat. Sci. **9**:307–319.

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

HAMMER, M. F., T. KARAFET, A. RASANAYAGAM, E. T. WOOD, T. K. ALTHEIDE, T. JENKINS, R. C. GRIFFITHS, A. R. TEMPLETON, and S. L. ZEGURA. 1998. Out of Africa and back again: nested cladistic analysis of human Y chromosome variation. Mol. Biol. Evol. **15**:427–441.

HARDING, R. M., S. M. FULLERTON, R. C. GRIFFITHS, J. BOND, M. J. COX, J. A. SCHNEIDER, D. S. MOULIN, and J. B. CLEGG. 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, and S. T. SHERRY. 1998. Genetic traces of ancient demography. Proc. Natl. Acad. Sci. USA **95**:1961–1967.

HEYER, E., J. PUYMIRAT, P. DIELTJES, E. BAKKER, and P. DE KNIJFF. 1997. Estimating Y chromosome specific microsatellite mutation frequencies using deep-rooting pedigrees. Hum. Mol. Genet. **6**:799–803.

HUDSON, R. R. 1990. Gene genealogies and the coalescent process. Pp. 1–44 *in* D. FUTUYMA and J. ANTONOVICS, eds. Oxford surveys in evolutionary biology. Vol. 7. Oxford University Press, Oxford, England.

JORDE, L. B., M. J. BAMSHAD, W. S. WATKINS, R. ZENGER, A. E. FRALEY, P. A. KRAKOWIAK, K. D. CARPENTER, H. SOODYALL, T. JENKINS, and A. R. ROGERS. 1995. Origins and affinities of modern humans: a comparison of mitochondrial and nuclear genetic data. Am. J. Hum. Genet. **57**:523–538.

KIMMEL, M., R. CHAKRABORTY, J. P. KING, M. BAMSHAD, W. S. WATKINS, and L. B. JORDE. 1998. Signatures of population expansion in microsatellite repeat data. Genetics **148**: 1921–1930.

MERRIWETHER, D. A., A. G. CLARK, S. W. BALLINGER, T. G. SCHURR, H. SOODYALL, T. JENKINS, S. T. SHERRY, and D. C. WALLACE. 1991. The structure of human mitochondrial DNA variation. J. Mol. Evol. **33**:543–555.

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.

PEREZ-LEZAUN, A., F. CALAFELL, M. SEIELSTAD, E. MATEU, D. COMAS, E. BOSCH, and J. BERTRANPETIT. 1997. Population genetics of Y-chromosome short tandem repeats in humans. J. Mol. Evol. **45**:265–270.

PRITCHARD, J. K., and M. W. FELDMAN. 1996. Statistics for microsatellite variation based on coalescence. Theor. Popul. Biol. **45**:265–270.

SEIELSTAD, M. T., E. MINCH, and L. L. C. CAVALLI-SFORZA. 1998. Genetic evidence for a higher female migration rate in humans. Nat. Genet. **20**:278–280.

SLATKIN, M. 1995. A measure of population subdivision based on microsatellite allele frequencies. Genetics **139**:457–462.

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

STONEKING, M., J. J. FONTIUS, S. L. CLIFFORD, H. SOODYALL, S. S. ARCOT, N. SAHA, T. JENKINS, M. A. TAHIR, P. L. DEININGER, and M. A. BATZER. 1997. Alu insertion polymorphisms and human evolution: evidence for larger population size in Africa. Genome Res. **7**:1061–1070.

TAVARÉ, S., D. J. BALDING, R. C. GRIFFITHS, and P. DONNELLY. 1997. Inferring coalescence times from DNA sequence data. Genetics **145**:505–518.

UNDERHILL, P. A., L. JIN, A. A. LIN, S. Q. MEHDI, T. JENKINS, D. VOLLRATH, R. W. DAVIS, L. L. CAVALLI-SFORZA, and P. J. OEFNER. 1997. Detection of numerous Y chromosome biallelic polymorphisms by denaturing high-performance liquid chromatography. Genome Res. **7**:996–1005.

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.

WEBER, J. L., and C. WONG. 1993. Mutation of human short tandem repeats. Hum. Mol. Genet. **2**:1123–1128.

WEISS, G., and A. VON HAESELER. 1996. Investigating the age of the common ancestor of men from the ZFY intron. Science **272**:1359–1360.

———. 1998. Inference of population history using a likelihood approach. Genetics **149**:1539–1546.

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., and D. J. BALDING. 1998. Genealogical inference from microsatellite data. Genetics **150**:499–510.

CHARLES F. AQUADRO, reviewing editor

Accepted September 2, 1999