Demographic estimates from Y chromosome microsatellite polymorphisms: Analysis of a worldwide sample

J. Michael Macpherson,[1] Sohini Ramachandran,[1] Lisa Diamond[1,2] and Marcus W. Feldman[1]*

[1]Department of Biological Sciences, Stanford University, Stanford, CA 94305, USA
[2]Stanford Genome Technology Center, 855 California Avenue, Palo Alto, CA 94304, USA
*Correspondence to: Tel: + 1 650 723 4952; Fax: + 1 650 725 0180; E-mail: macpher@stanford.edu

Date received (in revised form): 2nd July 2004

Abstract

Polymorphisms in microsatellites on the human Y chromosome have been used to estimate important demographic parameters of human history. We compare two coalescent-based statistical methods that give estimates for a number of demographic parameters using the seven Y chromosome polymorphisms in the HGDP-CEPH Cell Line Panel, a collection of samples from 52 worldwide populations. The estimates for the time to the most recent common ancestor vary according to the method used and the assumptions about the prior distributions of model parameters, but are generally consistent with other global Y chromosome studies. We explore the sensitivity of these results to assumptions about the prior distributions and the evolutionary models themselves.

Keywords: *Bayesian inference, demographic parameters, human history, statistical genetics, time to the most recent common ancestor (TMRCA), Y chromosome*

Introduction

It is possible to estimate evolutionary and demographic parameters from observed genetic variation in contemporary human populations. Of these quantities, the mean time to the most recent common ancestor (TMRCA) of the sample is of particular interest in population genetic studies because the age of the common ancestor indicates the relatedness of the individuals sampled.

Studies of coalescence in human DNA sequences usually focus on the uniparentally inherited Y chromosome and mitochondrial genome, where the absence of recombination limits the complexity of genealogical analysis.

The human Y chromosome is non-recombining over most of its length and is thus a highly informative haplotypic system that permits the tracing of paternal lineages and complements the maternal history of a population, as observed from mitochondrial DNA. Earlier studies have observed a high degree of geographic population structure on the Y chromosome, attributed to mating practices and the small effective population size of the Y chromosome.[1–4] Analyses of Y chromosomal haplotypes have been used to investigate the origins of specific regional populations. For example, studies have considered Austronesian-speaking populations,[5] histories of males in Israeli and Palestinian Arab populations,[6] and the history of Khoisan languages characterised by click consonants.[7] Fewer studies have looked at Y chromosome markers in globally distributed populations to calculate a TMRCA.[8–12] In this study, we investigate the global TMRCA, ancestral population size, growth rate and mutation rate from Y chromosome microsatellite polymorphisms in the HGDP-CEPH Human Genome Diversity Cell Line Panel.[13]

We contrast two coalescent-based methods of inference: (1) a modified version[9] of a rejection algorithm (RA)[14] and (2) the Markov Chain Monte Carlo (MCMC) program BATWING.[12] Both methods aim to produce posterior distributions for each of the above parameters, given a particular set of prior distributions on their values. These posterior distributions can differ between the two methods for a given set of priors and also are sensitive to the particular prior set chosen. We investigate both this sensitivity to choice of priors and the robustness of the inferences to changes in the underlying models.

Methods

Data

The 677 males in the sample come from 52 populations in seven geographical regions (Africa, America, Central/South Asia, East Asia, Europe, the Middle East and Oceania). The individuals were typed at seven polymorphic microsatellite loci on the non-recombining portion of the Y chromosome. These loci include two trinucleotide repeats (DYS388 and DYS392) and five tetranucleotide repeats (DYS389a, DYS389b, DYS390, DYS391 and DYS395). Across all loci, the sample contains 50 alleles, six of which are found only in a single population.

Computational methods

Our goal was to infer the joint distribution of several demographic and genetic parameters, given the polymorphism data. To do this, we used two methods, RA[9,14] and BATWING, a MCMC implementation.[12,15] Both methods assume the same growth model, in which the population has a constant effective number of Y chromosomes, N_A, until a time t_0 before the present. After this time, the population grows exponentially at a rate r_0 per generation. Each method uses these parameters together with the coalescent process[16,17] to generate genealogical trees with appropriately scaled branch lengths. Both also assume that mutations occur independently at each locus as a Poisson process with rate μ, which has as units mutations per locus per generation. Both methods require that prior distributions be specified for each of the previously mentioned parameters. As employed, neither method takes into account the possible effects of recombination, selection or population structure.

The key difference between the methods is that the RA uses summary statistics, while BATWING uses the full data. For this reason, the RA runs much faster than BATWING. In the RA, a genealogy is simulated under a parameter set sampled independently from the priors. If the standardised differences between each of three summary statistics computed for the simulated data and the observed data are all smaller than a threshold δ, the parameter set is accepted into the posterior distribution; if not, the parameter set is rejected. After many repetitions of this process, the collection of accepted sets of parameters forms the joint posterior distribution. The three summary statistics — number of haplotypes, mean variance in repeat number and mean heterozygosity — were chosen for their sensitivity to changes in population size.[9] Beaumont *et al.*[18] investigated the effects of using more summary statistics and a more sophisticated criterion for acceptance–rejection and found that results differed little from the approach of Pritchard *et al.*,[9] provided that the acceptance threshold was set low enough. We used a threshold of $\delta = 0.1$ in runs of at least 1 million trials, which normally resulted in acceptance rates of around 10^{-3}, well within the range recommended in Beaumont *et al.*[18]

While the RA starts afresh with each iteration, BATWING maintains a tree at all times and progresses by proposing new trees slightly different from the current tree. A new tree replaces the current tree probabilistically. By construction, BATWING's probabilities of transition between trees specify an irreducible Markov chain which is guaranteed to converge upon the joint distribution of interest, although there is no bound on the length of time convergence may require.[15] The computational expense in generating this potentially enormous number of iterations is BATWING's primary limitation (see 'Discussion on the paper' section of Wilson *et al.*[12]).

When BATWING simulates a mutation event, it is assumed that the number of microsatellite repeats changes by exactly one, with equal probability of increasing or decreasing. This, the stepwise mutation model (SMM),[19] is also the default model used by the RA. We experimented with two other mutation models using the RA, namely the symmetric geometric model (SGM),[9] in which the number of repeats changes by a value chosen from a symmetric geometric distribution having variance σ^2, and the range-constraint model (RCM),[20] in which the repeat number has stepwise changes but with hard reflecting boundaries located at a fixed number of repeats on either side of the original value. We set this fixed number to three, leading to a range of six, because the mean observed range of the number of repeats was 5.9.

We used four different sets of priors labelled *P*, *K*, *W* and *Z*, each consisting of a density function for each of the four parameters under both BATWING and the RA (Table 1). *P* and *W* derive from two previous global TMRCA studies of Y chromosome microsatellites, those of Pritchard *et al.*[9] and Wilson *et al.*[12] *K* and *Z* contrast higher and lower mutation rate means, as reported in the recent literature in Kayser *et al.*[21] and Zhivotovsky *et al.*[22] *P* and *W* use very diffuse priors for N_A, while *K* and *Z* use priors with mean N_A equal to 1,000, the value strongly suggested by Pritchard *et al.*[9] The respective priors for r_0 and t_0 are diffuse and identical across the four prior sets.

About 6 per cent of the repeat scores (290 of 4,739) are missing from the dataset. Because the number of haplotypes is not defined when data are missing, these missing data must be removed or replaced to allow the RA to proceed. Since most haplotypes which had missing data only lacked a single repeat score, we replaced each missing repeat score at a given locus by a value chosen from a multinomial distribution created from the frequencies of repeat scores observed for the respective population sample at that locus. Although BATWING can handle missing data by treating missing leaves as nuisance parameters, for consistency one such substituted dataset served for all the results reported here. We found that BATWING runs on the unsubstituted data resulted in posteriors very similar to those with the substituted data (results not shown). The dataset used in this analysis can be found at http://charles.stanford.edu/datasets.html.

Table 1. Prior distributions used to analyse microsatellite polymorphisms on the Y chromosome. The rejection algorithm[9] and BATWING[12] were run on each set of priors. Distributions were chosen based on past studies; the means for each distribution are given in brackets. μ is the mutation rate per locus per generation, N_A is the ancestral population size, t_0 is the time of start of exponential population growth in generations before present, r is rate of exponential population growth per generation

Prior set	Derived from	μ prior [mean]	N_A prior [mean]	t_0 prior [mean]	r prior [mean]
P	Pritchard *et al. (1999)*	gamma (10, 12,500) [0.0008]	Log normal (8.5, 2) [36,000]	exp (0.001) [1000]	gamma (1, 200) [0.005]
K	Kayser *et al. (2000)*	gamma (1, 416) [0.0024]	gamma (3, 0.003) [1000]	exp (0.001) [1000]	gamma (2, 400) [0.005]
W	Wilson, Weale and Balding (2003)	gamma (18, 8,170) [0.0022]	gamma (3, 0.001) [3000]	exp (0.001) [1000]	gamma (2, 400) [0.005]
Z	Zhivotovsky *et al. (2004)*	gamma (1.5, 2,175) [0.0069]	gamma (3, 0.003) [1000]	exp (0.001) [1000]	gamma (2, 400) [0.005]

Since our focus in this study was on the properties of the posterior distribution for the TMRCA and the demographic parameters, rather than the branching pattern of the genealogy, we modified the BATWING source code to reduce computation time. Reasoning that a maximum parsimony tree would have an approximately correct topology, we started BATWING with such a tree and disabled branch swapping after the first 100,000 iterations, which reduced the time of each iteration thereafter by about 30 per cent.

Results

Heterozygosity was computed for each of the seven major geographical regions using an unbiased estimator.[23] These ranged as follows: 0.45 (America), 0.53 (Africa), 0.55 (Middle East), 0.59 (Europe), 0.60 (Oceania), 0.62 (Central/South Asia) and 0.66 (East Asia). We performed an analysis of molecular variance using Genetic Data Analysis[24] and observed that 73–89 per cent of genetic variation occurs between individuals in the same population (Table 2). The American and Middle Eastern populations have especially low within-population variance. Of the six alleles found only in a single population, four appeared exactly once in the dataset, while the other two appeared twice. There were 46 alleles appearing more than once in the sample, of which three were exclusive to one of the seven major geographical regions listed in Table 2.

An important feature of human population genetic structure is the fraction of the total genetic variation that lies within populations relative to that among populations.[25–28] Due to the lower effective population sizes of the X and Y chromosomes and the mitochondrial genome compared with the autosomes, genetic drift is stronger in these systems, which can explain the smaller within-group variance that has been reported in such nonautosomal regions.[27,29] A correction suggested by Pérez-Lezaun *et al.*[30] can be applied. In the fifth column of Table 2, we see that the within-population components, after correction for population size, are generally similar to those reported by Rosenberg *et al.*[28] and Ramachandran *et al.*[29] When this analysis is repeated using only the tetranucleotide repeats, the results are not affected.[31,32]

We have summarised the posterior distributions from the RA and BATWING for each of the four sets of prior distributions in Table 3. The three sets of priors *P*, *K* and *Z* tend to produce similar posteriors under the RA and BATWING, including TMRCA point estimates of 60,000–90,000 years before the present (ybp), assuming a constant generation length of 25 years. Note that BATWING may only be compared directly to the RA using the SMM. In many cases, as would be expected from its use of the full data, BATWING produced narrower credible intervals than did the RA, but this reduction in the variance was not universal. It can be seen in Figure 1 that for the *P*, *K* and *Z* priors, the TMRCA traces which form the BATWING posteriors have most support in the region from 60,000–90,000 ybp, and differ in the extent to which the priors permit exploration of parameter space.

BATWING and the RA gave different point estimates for t_0, r, and N_A. The RA estimated a growth period beginning 20,000–25,000 ybp, growing at a rate of $6-8 \times 10^{-3}$ per generation, whereas BATWING placed greatest support on a longer, slower growth period of 30,000–50,000 years at rate of $3-5 \times 10^{-3}$ per generation. BATWING also tended to give smaller estimates of ancestral population size than the RA (mean 700–1,000, compared with 1,000–1,500). The mutation rate estimates were similar with both approaches, at $7-9 \times 10^{-4}$ per locus per generation. Despite the differences in the modes of the distributions, these posterior distributions overlap considerably for each demographic parameter.

The use of *W* priors resulted in posteriors very different from those of the *P*, *K* and *Z* priors, namely a much younger TMRCA point estimate of 30,000 ybp, and a much greater mutation rate and growth rate (Table 3). We address the discrepancy between the results of the *W* priors and the other three prior sets in the Discussion.

Table 2. Analysis of molecular variance for the Y chromosome. Ninety-five per cent confidence intervals (CIs; in parentheses) were calculated using 1,000 bootstraps across loci. The World-B97 sample[28] consists of 14 populations that were chosen in order to approximate the sample of Barbujani *et al.*[26] The fifth column corrects for the smaller Y chromosome population size, as in Pérez-Lezaun *et al.*[30] The estimate and CI for the among-region variance component for Eurasia are set to zero because Weir's unbiased estimator[23] yields slightly negative values

Sample	Number of regions	Number of populations	Variance components and 95% confidence intervals (%)			
			Within populations	Within populations (corrected)	Among populations within regions	Among regions
World	1	52	80.4 (74.7–84.5)	94.3 (92.2–95.6)	19.6 (15.5–25.2)	
World	5	52	80.2 (73.2–85.6)	94.2 (91.6–96.0)	15.0 (13.2–17.1)	4.80 (0.00–10.6)
World	7	52	83.9 (77.5–88.8)	95.4 (93.2–96.9)	15.5 (11.2–17.2)	0.56 (0.00–6.16)
World-B97	5	14	84.7 (71.5–97.1)	95.7 (90.9–99.3)	8.43 (2.90–11.1)	6.85 (0.00–22.1)
Africa	1	6	91.2 (87.6–94.5)	97.6 (96.6–98.6)	8.78 (5.49–12.4)	
Eurasia	1	21	86.4 (83.4–89.1)	96.2 (95.3–97.0)	13.6 (10.9–16.6)	
Eurasia	3	21	88.9 (85.4–91.7)	97.0 (95.9–97.8)	11.1 (8.25–14.4)	0.00 (0.00–0.00)
Europe	1	8	86.8 (81.3–92.8)	96.3 (94.6–98.1)	13.2 (7.17–18.7)	
Middle East	1	4	66.2 (58.7–74.6)	88.7 (85.0–92.2)	33.8 (25.4–41.2)	
Central/ South Asia	1	9	94.7 (92.0–97.2)	98.6 (97.9–99.3)	5.30 (2.77–8.04)	
East Asia	1	18	81.0 (71.8–87.7)	94.5 (91.1–96.6)	19.0 (12.3–28.2)	
Oceania	1	2	80.6 (70.2–92.3)	94.3 (90.4–98.0)	19.4 (7.65–29.8)	
America	1	5	58.0 (48.2–70.0)	84.7 (78.8–90.3)	42.0 (29.9–51.7)	

Effect of mutation and growth models

Because we were interested in the effects of model assumptions on the posterior distributions, we tested three models of mutation and growth using the RA. Following Pritchard *et al.*,[9] we used the relative acceptance rate ratios to determine which model, if any, was more consistent with the data. This amounts to placing an evenly weighted prior on each model and using the RA to assess posterior support for the models.

We observed that the TMRCA posteriors and, to a lesser extent, those of the ancestral population size and the mutation rate, were affected by the choice of mutation model (Table 3). The RCM produced considerably longer TMRCAs (80,000–120,000 ybp) than the SMM, while the SGM produced shorter TMRCAs (30,000–70,000 ybp). This ordering matches that of Pritchard *et al.*,[9] who used the RA with a smaller dataset of similar geographical diversity. The acceptance rates for all three models were similar (Table 4).

Since some studies have documented asymmetry in the microsatellite mutation process (eg Calabrese and Durrett[33]), we compared the RA to a version adapted to permit mutational asymmetry. Using the SGM and the *K* priors, identical and independent priors were placed on the respective rates of repeat increase and decrease. The mean acceptance rates were nearly identical for the symmetric and asymmetric mutation models (symmetric/asymmetric ratio 52/48; mean acceptance rate 1.63×10^{-3}). The posteriors for the increase and decrease rates were very similar and also very similar to the μ posterior for the mutation model. Thus, these data are not better explained by the asymmetric model than the symmetric mutation model.

Several studies[21,34] have noted that microsatellite loci on the same chromosome can mutate at different rates. We addressed the assumption that each locus mutates at the same rate by comparing RA acceptance rates between runs assuming that all loci mutate at the same rate and runs assuming that each locus mutates at a different rate. We used the *Z* priors in both cases; in the latter case, one mutation rate was drawn from the *Z* μ prior for each locus. The acceptance rate was similar to that obtained when a single mutation rate governed all the loci (single rate/multiple rates ratio 49/51; mean acceptance rate 1.42×10^{-3}), as were the estimated TMRCAs and parameter posteriors. Both this result and the mutation rate symmetry result held for all sets of priors.

Table 3. Demographic parameters estimated from seven Y chromosome microsatellite loci in 677 individuals drawn from 52 global populations. Columns: priors (see Table 1 for details of each set of prior distributions); method (RA = rejection algorithm using stepwise mutation model [SMM], symmetrical geometric model [SGM], or range constraint model [RCM], BW = BATWING followed by number of Markov Chain Monte Carlo updates); remaining columns are posterior distributions, with mean and 95 per cent credible interval, TMRCA is the mean time to the most recent common ancestor in years before present (generation time is assumed to be 25 years), μ is mutation rate per locus per generation, N_A is ancestral population size, t_0 is time of start of exponential population growth in generation before present, r is rate of exponential population growth per generation

Prior set	Method	TMRCA ($\times 10^{-3}$) mean [95%]	μ ($\times 10^{-3}$) mean [95%]	N_A mean [95%]	t_0 ($\times 10^{3}$) mean [95%]	r ($\times 10^{-2}$) mean [95%]
P	RA, SMM	86 [30–221]	0.71 [0.34–1.19]	1630 [140–4520]	22 [8–51]	0.77 [0.23–2.13]
P	RA, SGM	58 [20–147]	0.71 [0.36–1.18]	990 [70–3110]	23 [8–52]	0.75 [0.25–2.08]
P	RA, RCM	121 [33–441]	0.73 [0.35–1.22]	2070 [240–6510]	21 [8–50]	0.79 [0.22–2.16]
P	BW, 4×10^6	82 [34–195]	0.80 [0.40–1.34]	2510 [400–11630]	36 [12–73]	0.32 [0.15–0.58]
P	BW, 200×10^6	64 [31–131]	0.79 [0.39–1.32]	710 [250–1740]	48 [24–92]	0.30 [0.14–0.52]
K	RA, SMM	68 [21–178]	0.89 [0.27–2.15]	1020 [220–2410]	24 [7–64]	0.67 [0.20–1.55]
K	RA, SGM	64 [21–155]	0.66 [0.20–1.59]	960 [220–2270]	27 [9–61]	0.67 [0.21–1.60]
K	RA, RCM	78 [22–199]	0.98 [0.29–2.51]	1090 [290–2390]	22 [7–55]	0.69 [0.23–1.55]
K	BW, 4×10^6	55 [16–143]	1.27 [0.31–4.02]	760 [190–2350]	42 [4–118]	0.43 [0.11–1.33]
K	BW, 200×10^6	63 [17–178]	1.10 [0.31–2.71]	660 [190–1610]	52 [13–159]	0.37 [0.11–0.92]
K	BW, 800×10^6	61 [18–164]	0.90 [0.27–2.12]	760 [230–1760]	42 [11–124]	0.47 [0.14–1.10]
W	RA, SMM	39 [19–81]	1.68 [1.01–2.72]	940 [250–1940]	11 [5–18]	1.01 [0.46–1.82]
W	RA, SGM	29 [14–57]	1.49 [0.91–2.18]	600 [230–1300]	13 [7–21]	0.85 [0.49–1.29]
W	RA, RCM	83 [20–266]	1.76 [1.08–2.71]	1470 [250–3730]	11 [5–24]	0.87 [0.36–1.95]
W	BW, 4×10^6	32 [16–66]	1.85 [1.10–2.78]	1100 [260–4160]	14 [5–26]	0.72 [0.41–1.17]
W	BW, 200×10^6	29 [15–57]	1.81 [1.07–2.74]	590 [240–1210]	16 [9–27]	0.70 [0.40–1.09]
Z	RA, SMM	79 [27–200]	0.71 [0.24–1.54]	1120 [290–2550]	26 [9–63]	0.62 [0.20–1.48]
Z	RA, SGM	71 [24–170]	0.55 [0.18–1.24]	1010 [250–2310]	30 [10–74]	0.62 [0.20–1.42]
Z	RA, RCM	88 [28–215]	0.74 [0.25–1.61]	1210 [330–2660]	26 [9–65]	0.64 [0.20–1.49]
Z	BW, 4×10^6	41 [16–94]	1.40 [0.52–2.94]	1490 [540–3420]	23 [5–60]	0.48 [0.18–1.04]
Z	BW, 200×10^6	84 [26–228]	0.72 [0.22–1.64]	690 [210–1630]	80 [22–232]	0.25 [0.08–0.57]

We compared the growth model described above with a model having two exponential growth phases using a modified version of the RA. This second model has additional parameters t_1, the time before the present at which the population begins its second growth phase, and r_1, the rate at which the population grows thereafter. Over the worldwide data, acceptance rates were similar between the two models (single growth-phase/dual growth-phase ratio 52/48; mean acceptance rate 1.88×10^{-3}). Under the dual growth phase model using the *P* priors, growth was somewhat slower and lasted longer than in the single growth phase model (Table 5). This result was also robust to the choice of prior set.

We also compared the single growth model against a model with constant population size. In agreement with Pritchard *et al.*,[9] the growth model had a much higher acceptance rate (constant/growth ratio 0/100; mean acceptance rate 0.68×10^{-3}) than did the model of constant population size, with the exception of two populations, namely the Oceanic and American populations, each of which had slightly higher acceptance rates for the constant model than for the expansion model (see also Zhivotovsky *et al.*[35]).

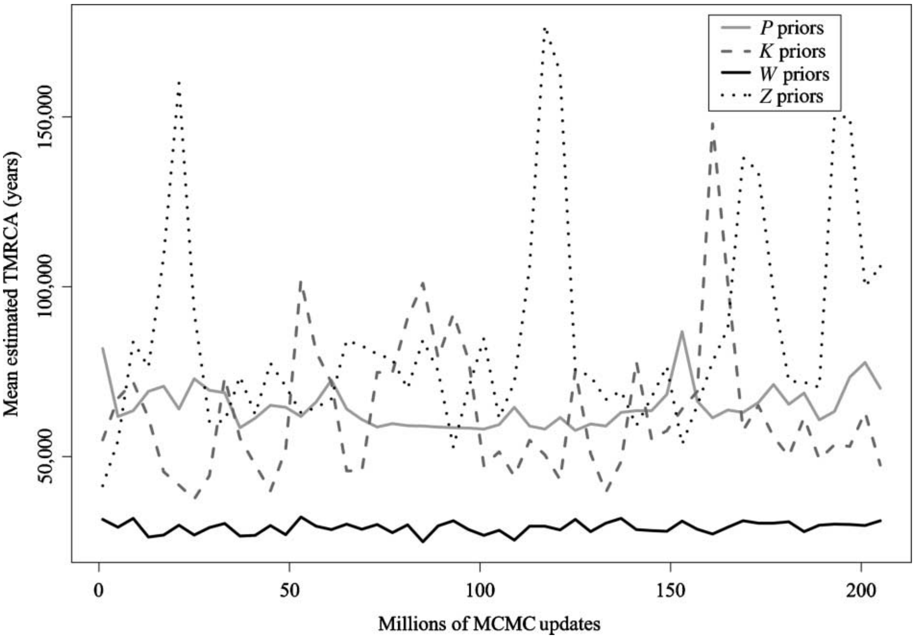

Figure 1. The most recent common ancestor (TMRCA) estimated from BATWING runs over the four sets of priors. Each plot point is the mean of 4 million Markov Chain Monte Carlo (MCMC) updates

Using regional subsets of the data, both methods produced different posteriors from those using the entire sample. Table 6 shows results obtained with the RA using the *P* priors with the SGM from typical runs of both the RA and BATWING. There is great overlap between these posteriors, and in the cases of the TMRCA and the time of expansion, this overlap makes it difficult to discern a clear pattern in the timing of splitting of the population or expansion of the respective subpopulations.

With such a large dataset, BATWING required a long time to converge. We observed that after long runs of 200 million iterations, the posteriors that resulted differed from those produced after the first 4 million iterations (Table 3). We monitored progress towards convergence by computing the autocorrelation function (ACF), the correlation of a chain with itself when its indices are offset by some integer lag, and observing the chains and the overall likelihood. In BATWING runs using the entire dataset, we found that the ACF decayed to zero monotonically but slowly as the lag increased. We extended several BATWING runs to hundreds of millions of iterations and watched for signs of nonconvergence, but found that the parameter plots and likelihood plots remained steady (Figure 1). The ACF continued to decline monotonically and more rapidly than before the runs were extended, but it still did not reach zero for lags of less than tens of thousands for any of the demographic parameter chains. This differs from the

Table 4. Ratio of acceptances per million trials at threshold $\delta = 0.1$ and average acceptance rates using the rejection algorithm[9] and different mutation models for all four sets of priors (Table 1)

Prior set	Mutation model			
	SGM/ SMM	SMM/ RCM	SGM/ RCM	Mean rate ($\times 10^{-3}$)
P	39/61	44/56	33/67	1.34
K	43/57	49/51	42/58	1.47
W	16/84	30/70	8/92	0.14
Z	52/48	49/51	48/52	2.82

Table 5. Results from rejection algorithm with one and two exponential growth phases using *P* priors

	(Mean, 95% range)			
	Single-phase		Dual-phase	
T	58,000	(20,000, 14,7000)	46,000	(18,000, 130,000)
N_A	990	(44, 3,100)	880	(55, 3,400)
r_0	0.0075	(0.0025, 0.028)	0.0049	(0.00018, 0.015)
r_1	–		0.0062	(0.00028, 0.018)
t_0	23,000	(8000, 52,000)	12,000	(620, 32,000)
t_1	–		13,000	(430, 41,000)

Table 6. Mean parameter estimates and 95 per cent credible intervals for individual populations obtained with the rejection algorithm using the *P* priors and the symmetrical geometric model

Population	TMRCA ($\times 10^3$) mean [95%]	μ ($\times 10^{-3}$) mean [95%]	N_A mean [95%]	t_0 ($\times 10^3$) mean [95%]	R ($\times 10^{-2}$) mean [95%]
World	58 [20–150]	0.71 [0.34–1.19]	990 [70–3100]	22 [8.5–50]	0.83 [0.26–2.0]
Africa	53 [15–160]	0.70 [0.35–1.25]	1,000 [62–3500]	16 [3.8–43]	0.64 [0.12–1.8]
Non-Africa	54 [20–140]	0.70 [0.34–1.17]	970 [64–3500]	23 [8.4–51]	0.81 [0.26–2.1]
America	30 [11–83]	0.74 [0.35–1.25]	730 [57–2100]	14 [1.0–48]	0.43 [0.18–1.7]
Central/South Asia	51 [18–140]	0.73 [0.34–1.27]	970 [61–3300]	21 [6.8–48]	0.65 [0.19–1.8]
East Asia	55 [20–150]	0.72 [0.36–1.26]	1,000 [59–3500]	23 [7.8–53]	0.70 [0.21–1.9]
Europe	44 [16–110]	0.69 [0.34–1.19]	800 [52–2700]	20 [7.0–48]	0.70 [0.21–1.9]
Eurasia	51 [19–140]	0.70 [0.35–1.21]	940 [92–3100]	20 [7.0–48]	0.81 [0.25–2.1]
Middle East	52 [15–150]	0.76 [0.37–1.33]	1,100 [96–3300]	15 [1.7–48]	0.43 [0.31–1.3]
Oceania	59 [18–160]	0.76 [0.36–1.32]	1,400 [120–4100]	17 [0.8–55]	0.37 [0.15–1.4]

experience of Wilson *et al.*,[12] who reported ACF declining to zero by lag 30. It is likely that this difference is caused by slow movement of the chain between distant regions of parameter space, which might be expected, since the number of nuisance parameters — for example the internal node haplotypes — and therefore the size of the parameter space, is much larger in this study than in Wilson *et al.*[12]

Discussion

These two methods of inference support a recent human Y chromosome TMRCA. Three of the sets of priors we examined resulted in a mean TMRCA of 60,000–90,000 ybp. The estimates exceeded 100,000 ybp when we limited the range of mutation under the RA. These values are consistent with other global TMRCA estimates from Y chromosome microsatellite data, including those of Pritchard *et al.*[9] (46,000–91,000 ybp), and also concur with several single nucleotide polymorphism studies of the Y chromosome, including Thomson *et al.*[10] (48,000–59,000 ybp) and Tang *et al.*[36] (91,000 ybp). Interpreting these estimates requires care, because they are sensitive to both the priors and models assumed, and rely on a simple model of evolution.

Sensitivity to priors

We examined the dependence of the results from the two methods on the different sets of priors. It is clear from Table 3, particularly in the case of the *W* priors, that the choice of priors affects the posteriors. There are two main reasons for the posteriors not to be identical for different sets of priors. First, the data may not be informative. Uninformative data would imply a flat likelihood surface. We would expect to see posteriors resembling the priors under both methods. The microsatellite data used here appear to be informative, because the posteriors differ from their priors (Figure 2), and because the RA and BATWING tended to infer similar posterior departures from a given set of priors (Table 3). The second reason may be that the priors do not allow the exploration of those regions of the parameter space which would otherwise be included in the posteriors; this appears to explain the divergent results obtained with the *W* priors.

Using the *W* priors from Wilson *et al.*[12] with the HGDP-CEPH data, both inference methods produced a mean TMRCA estimate of around 30,000 years, which is consistent with the findings of that paper. The *W* μ prior has much smaller variance than the *K* μ prior, effectively precluding values of μ smaller than 0.001 (Figure 2). The *K* and *Z* μ priors include these smaller values as well as the higher values implied under the *W* priors. With both low and high mutation rates available, the posteriors from the *K* and *Z* priors placed most of their support on mutation rates lower than 0.001 (Figure 2), leading to an older TMRCA. Because of this result, and since a TMRCA of 30,000 years seems improbable in light of archaeological evidence that anatomically modern humans existed outside Africa at least 35,000–45,000 ybp,[37–40] we suggest that lower mutation rates leading to an earlier TMRCA are more plausible than the higher rates of the *W* priors.

The inadvertent restriction of parameter space might be mitigated by choosing uniform priors spanning the conceivable range of the parameters. For large datasets, this is not a feasible approach for the RA, much less for BATWING, with available computers. A practicable approach might be to use the RA to compute the standardised differences between the summary statistics and the data as usual, and to simply record these

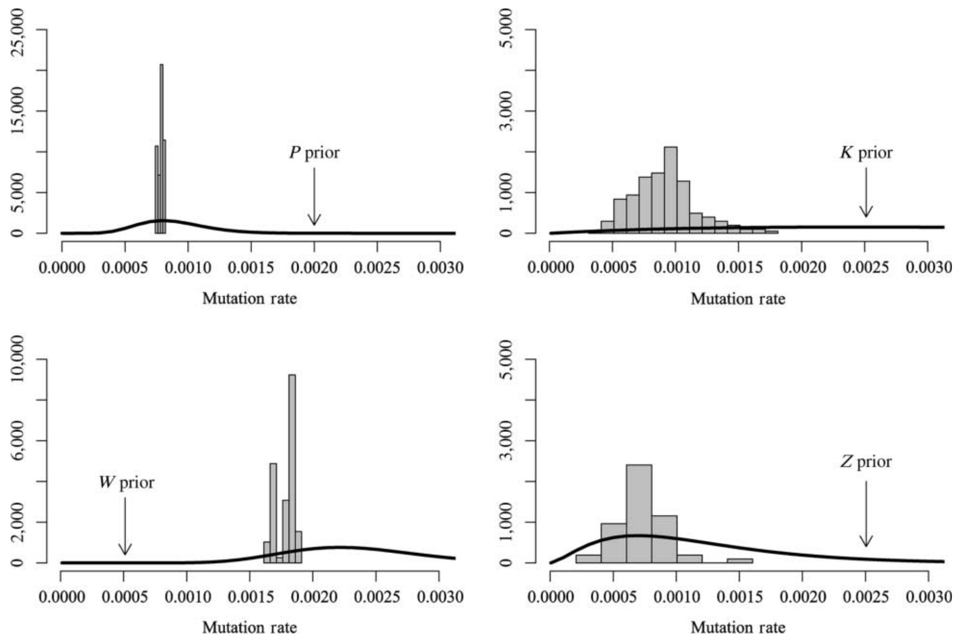

Figure 2. Comparison of mutation rate prior distributions and posterior distributions estimated by BATWING between four sets of priors. Priors are represented by curves, posteriors by histograms

differences rather than accepting or rejecting them, drawing the parameters from broad uniform priors. Some regions of parameter space will likely generate very different summary statistics from the data, and these may be ignored. But those parameter space regions which produce statistics near to those of the data may serve as a basis for establishing sensible priors.

Y chromosome mutation rates

A phylogenetic study by Zhivotovsky *et al.*[23] estimated the mean effective Y chromosome mutation rate to be $0.69 \times 10^{-3} \pm 0.57 \times 10^{-3}$. By contrast, estimates from pedigree studies suggest higher mutation rates (2.8×10^{-3} and 2.1×10^{-3})[21,41]. Two of our sets of prior distributions, K and Z, differed only in the priors for μ, with a higher mean of 2.4×10^{-3} and larger variance in the former, and a lower mean of 0.69×10^{-3} with smaller variance in the latter. The resulting posteriors, with respective means of 0.92×10^{-3} and 0.73×10^{-3}, tend to support an effective mutation rate more like the lower, phylogeny-derived value.

It appears that the data are consistent with all three mutation models. The RCM limits the repeat length explicitly by prohibiting repeat lengths too distant from the initial value. Neither the SMM nor the SGM place restrictions on the length, but the SMM approaches large repeat lengths more slowly than the SGM because its step size is limited to one. The more that repeat lengths are limited, the longer it takes to achieve some level of diversity in the population, which accounts for the observation that the oldest TMRCA occurs with the RCM while the youngest occurs with the SGM model. We also found that the results were not sensitive to the assumption of a single mutation rate across all loci or the assumption of symmetry in repeat length change.

The estimated TMRCA is critically dependent on prior assumptions about the rate of mutation. Assuming constant population size and a given sample size, the coalescent branch length expectations are directly proportional to the population size; doubling the population size doubles the expected length of each branch in the genealogy. The coalescent in a growing population may also be rescaled, such that all branch lengths change in the same proportion, although this requires more than a simple proportional change in the population size.[42] For example, by appropriately scaling the demographic priors and mutation rate, we obtained arbitrarily large or small TMRCAs using the HGDP-CEPH data under both RA and BATWING, with acceptance rates indistinguishable from those reported in Table 4. The differences in acceptance rates between the four prior sets reflect not the likelihood of the TMRCA given the data, but the likelihood of the combination of tree geometry and mutation rate. Changing the population history changes the relative lengths of the tree branches; some tree geometries are more consistent with the dataset's level of polymorphism than others. The TMRCAs we report are those most consistent with the range of mutation rates reported in the literature.

Were there distinct epochs of population growth?

The estimates in this study of 20,000–50,000 ybp for the time of population expansion long precede 10,000 ybp, the time around which agriculture is widely believed to have developed,[37,43,44] and at which the population would naturally be expected to increase. Several other studies also give estimates of expansion time much earlier than 10,000 ybp, including microsatellite studies of the Y chromosome (20,000 ybp, from Pritchard *et al.*[9]) and autosomes (35,000 ybp, from Zhivotovsky *et al.*[22]). Estimates of expansion time from extensive studies of nuclear autosomal sequences, such as 0–100,000 ybp[45] and 36,000–97,000 ybp,[46] also suggest an early start to population expansion. If population increase began long before the development of agriculture, something else, perhaps another behavioural change, may have precipitated this earlier expansion.

Reasoning that the emergence of agriculture might have drastically increased the rate of population growth, we compared the original RA to a version which explicitly allowed for two distinct growth phases. We did not observe much difference between the results for the two growth phases (Table 5), nor between those for the two growth phases combined and the original, single phase of growth. Furthermore, the acceptance rates were quite similar between the two growth models. We conclude that, although we observe a strong signal of growth by comparison to the constant population size model, no sharp increase in the rate of growth after its onset is evident from the data.

Both methods explored here make a number of simplifying assumptions. While recombination can probably be safely disregarded for these Y chromosome markers, the same cannot be said for the possible effects of selection, population structure and sampling error. Selection is known to mimic population growth,[45] compressing towards the present the portion of the genealogy in which it acts. Population structure may also strongly affect genealogies,[47] as can the pooling of samples from different populations[48,49] and uncorrected ascertainment bias.[50] Inferences drawn from other genomic regions and from more specific models will be useful in more accurately understanding human demographic history.

Our estimates for Y chromosome TMRCAs are again shorter than those obtained for the mitochondrion,[36,51] reinforcing interest in understanding the differences between male and female demography for early modern humans. Further work might include implementing models of range constraints with soft boundaries for microsatellites.[52,53] The RA might also be modified to include population subdivision, with different expansion times for different populations.

Acknowledgments

We thank Thomas Lebarbanchon for contributions to the rejection algorithm implementation. Richard Klein, Noah Rosenberg, Hua Tang, Lev Zhivotovsky and our reviewers provided insightful comments on the manuscript. This work was supported by NIH grant GM28428 to MWF. JMM is a Howard Hughes Medical Institute Predoctoral Fellow. SR is supported by an NDSEG fellowship.

References

1. Seielstad, M.T., Minch, E. and Cavalli-Sforza, L.L. (1998), 'Genetic evidence for a higher female migration rate in humans', *Nat. Genet.* Vol. 20, pp. 278–280.
2. Bertranpetit, J. (2000), 'Genome, diversity, and origins: The Y chromosome as a storyteller', *Proc. Natl. Acad. Sci. USA* Vol. 97, pp. 6927–6929.
3. Underhill, P.A., Shen, P.D., Lin, A.A. *et al.* (2000), 'Y chromosome sequence variation and the history of human populations', *Nat. Genet.* Vol. 26, pp. 358–361.
4. Hammer, M.F., Karafet, T.M., Redd, A.J. *et al.* (2001), 'Hierarchical patterns of global human Y chromosome diversity', *Mol. Biol. Evol.* Vol. 18, pp. 1189–1203.
5. Hurles, M.E., Nicholson, J., Bosch, E. *et al.* (2002), 'Y chromosomal evidence for the origins of Oceanic-speaking peoples', *Genetics* Vol. 160, pp. 289–303.
6. Nebel, A., Filon, D., Weiss, D.A. *et al.* (2000), 'High-resolution Y chromosome haplotypes of Israeli and Palestinian Arabs reveal geographic substructure and substantial overlap with haplotypes of Jews', *Hum. Genet.* Vol. 107, pp. 630–641.
7. Knight, A., Underhill, P.A., Mortensen, H.M. *et al.* (2003), 'African Y chromosome and mtDNA divergence provides insight into the history of click languages', *Curr. Biol.* Vol. 13, pp. 464–473.
8. Hammer, M.F., Karafet, T., Rasanayagam, A. *et al.* (1998), 'Out of Africa and back again: Nested cladistic analysis of human Y chromosome variation', *Mol. Biol. Evol.* Vol. 15, pp. 427–441.
9. Pritchard, J.K., Seielstad, M.T., Pérez-Lezaun, A. and Feldman, M.W. (1999), 'Population growth of human Y chromosomes: A study of Y chromosome microsatellites', *Mol. Biol. Evol.* Vol. 16, pp. 1791–1798.
10. Thomson, R., Pritchard, J.K., Shen, P., Oefner, P.J. and Feldman, M.W. (2000), 'Recent common ancestry of human Y chromosomes: Evidence from DNA sequence data', *Proc. Natl. Acad. Sci. USA* Vol. 97, pp. 7360–7365.
11. Tang, H., Siegmund, D.O., Shen, P. *et al.* (2001), 'Estimation of the time to the most recent common ancestor by tree-partition', *Am. J. Hum. Genet.* Vol. 69(Supplement), pp. 394.
12. Wilson, I.J., Weale, M.E. and Balding, D.J. (2003), 'Inferences from DNA data: Population histories, evolutionary processes and forensic match probabilities', *J. R. Stat. Soc. [Ser A]* Vol. 166, pp. 155–201.
13. Cann, H.M., de Toma, C., Cazes, L. *et al.* (2002), 'A human genome diversity cell line panel', *Science* Vol. 296, pp. 261–262.
14. Tavaré, S., Balding, D.J., Griffiths, R.C. and Donnelly, P. (1997), 'Inferring coalescence times from DNA sequence data', *Genetics* Vol. 145, pp. 505–518.
15. Wilson, I.J. and Balding, D.J. (1998), 'Genealogical inference from microsatellite data', *Genetics* Vol. 150, pp. 499–510.
16. Kingman, J.F.C. (1982), 'The coalescent', *Stochastic Process. Appl.* Vol. 13, pp. 235–248.
17. Nordborg, M. (2003), 'Coalescent theory', in Balding, D.J., Bishop, M., Cannings, C. Eds., *Handbook of Statistical Genetics*, 2nd edn. Wiley, Chichester, W. Sussex, UK, pp. 602–635.
18. Beaumont, M.A., Zhang, W. and Balding, D.J. (2002), 'Approximate Bayesin computation in population genetics', *Genetics* Vol. 162, pp. 2025–2035.
19. Ohta, T. and Kimura, M. (1973), 'A model of mutation appropriate to estimate the number of electrophoretically detectable alleles in a finite population', *Genet. Res.* Vol. 22, pp. 201–204.
20. Feldman, M.W., Bergman, A., Pollock, D.D. and Goldstein, D.B. (1997), 'Microsatellite genetic distances with range constraints: Analytic description and problems of estimation', *Genetics* Vol. 145, pp. 207–216.
21. Kayser, M., Roewer, L., Hedman, M. *et al.* (2000), 'Characteristics and frequency of germline mutations at microsatellite loci from the human Y chromosome, as revealed by direct observation in father/son pairs', *Am. J. Hum. Genet.* Vol. 66, pp. 1580–1588.
22. Zhivotovsky, L.A., Underhill, P.A., Ginnioglu, C. *et al.* (2004), 'The effective mutation rate at Y chromosome short tandem repeats, with application to human population-divergence time', *Am. J. Hum. Genet.* Vol. 74, pp. 50–61.

23. Weir, B. (1996), Genetic Data Analysis II, Sinauer Press, Sunderland, MA.
24. Lewis, P.O. and Zaykin, D.V. (2001), 'Genetic data analysis: Computer program for the analysis of allelic data. Version 1.0 (d16c)', http://lewis.eeb.uconn.edu/lewishome/software.html.
25. Lewontin, R.C. (1972), 'The apportionment of human diversity', *Evol. Biol.* Vol. 6, pp. 381–398.
26. Barbujani, G., Magagni, A., Minch, E. and Cavalli-Sforza, L.L. (1997), 'An apportionment of human DNA diversity', *Proc. Natl. Acad. Sci. USA* Vol. 94, pp. 4516–4519.
27. Jorde, L.B., Watkins, W.S., Bamshad, M.J. *et al.* (2000), 'The distribution of human genetic diversity: A comparison of mitochondrial, autosomal, and Y chromosome data', *Am. J. Hum. Genet.* Vol. 66, pp. 979–988.
28. Rosenberg, N.A., Pritchard, J.K., Weber, J.L. *et al.* (2002), 'Genetic structure of human populations', *Science* Vol. 298, pp. 2381–2385.
29. Ramachandran, S., Rosenberg, N.A., Zhivotovsky, L.A. and Feldman, M.W. (2004), 'Robustness of the inference of human population structure: A comparison of X-chromosomal and autosomal microsatellites', *Hum. Genom.* Vol. 1, pp. 87–97.
30. Pérez-Lezaun, A., Calafell, F., Seielstad, M. *et al.* (1997), 'Population genetics of Y-chromosome short tandem repeats in humans', *J. Mol. Evol.* Vol. 45, pp. 265–270.
31. Excoffier, L. and Hamilton, G. (2003), 'Comment on 'Genetic structure of human populations'', *Science* Vol. 300, p. 1877.
32. Rosenberg, N.A., Pritchard, J.K., Weber, J.L. *et al.* (2003), 'Response to Comment on 'Genetic structure of human populations'', *Science* Vol. 300, p. 1877.
33. Calabrese, P. and Durrett, R. (2003), 'Dinucleotide repeats in the Drosophila and human genomes have complex, length-dependent mutation processes', *Mol. Biol. Evol.* Vol. 20, pp. 715–725.
34. Bowcock, A.M., Ruiz-Linares, A., Tomforhde, J. *et al.* (1994), 'High-resolution of human evolutionary trees with polymorphic microsatellites', *Nature* Vol. 368, pp. 455–457.
35. Zhivotovsky, L.A., Rosenberg, N.A. and Feldman, M.W. (2003), 'Features of evolution and expansion of modern humans, inferred from genome-wide microsatellite markers', *Am. J. Hum. Genet.* Vol. 72, pp. 1171–1186.
36. Tang, H., Siegmund, D.O., Shen, P.D. *et al.* (2002), 'Frequentist estimation of coalescence times from nucleotide sequence data using a tree-based partition', *Genetics* Vol. 161, pp. 447–459.
37. Klein, R.G. (1999), *The Human Career,* University of Chicago Press, Chicago, IL.
38. Mellars, P. (1998), 'The fate of the Neanderthals', *Nature* Vol. 395, pp. 539–540.
39. O'Connell, J.F. and Allen, J. (1998), 'When did humans first arrive in greater Australia and why is it important to know?', *Evol. Anthropol.* Vol. 6, pp. 132–146.
40. Bowler, J.M., Johnston, H., Olley, J.M. *et al.* (2003), 'New ages for human occupation and climatic change at Lake Mungo, Australia', *Nature* Vol. 421, pp. 837–840.
41. Heyer, E., Puymirat, J., Dieltjes, P. *et al.* (1997), 'Estimating Y chromosome-specific microsatellite mutation frequencies using deep rooting pedigrees', *Hum. Mol. Genet.* Vol. 6, pp. 799–803.
42. Griffiths, R.C. and Tavaré, S. (1994), 'Sampling theory for neutral alleles in a varying environment', *Phil. Trans. R. Soc. Lond. B* Vol. 344, pp. 403–410.
43. Ammerman, A.J. and Cavalli-Sforza, L.L. (1984), *The Neolithic Transition and the Genetics of Populations in Europe*, Princeton University Press, Princeton, NJ.
44. Cavalli-Sforza, L.L. and Feldman, M.W. (2003), 'The application of molecular genetic approaches to the study of human evolution', *Nat. Genet.* Vol. 33, pp. 266–275.
45. Wall, J.D. and Przeworski, M. (2001), 'When did the human population size start increasing?', *Genetics* Vol. 155, pp. 1865–1874.
46. Pluzhnikov, A., Di Rienzo, A. and Hudson, R.R. (1999), 'Inferences about human demography based on multilocus analyses of noncoding sequences', *Genetics* Vol. 161, pp. 1209–1218.
47. Beaumont, M.A. (2004), 'Recent developments in genetic data analysis: What can they tell us about human demographic history?', *Heredity* Vol. 922, pp. 365–379.
48. Ptak, S. and Przeworski, M. (2002), 'Evidence for population growth in humans is confounded by fine-scale population structure', *Trends Genet.* Vol. 18, pp. 559–593.
49. Hammer, M.F., Blackmer, F., Garrigan, D. *et al.* (2003), 'Human population structure and its effects on sampling Y chromosome sequence variation', *Genetics* Vol. 164, pp. 1495–1509.
50. Wakeley, J., Nielsen, R., Liu-Cordero, S.N. and Ardlie, K. (2001), 'The discovery of single-nucleotide polymorphisms — and inferences about human demographic history', *Am. J. Hum. Genet.* Vol. 69, pp. 1332–1347.
51. Vigilant, L., Stoneking, M., Harpending, H. *et al.* (1991), 'African populations and the evolution of human mitochondrial DNA', *Science* Vol. 253, pp. 1503–1507.
52. Garza, J.C., Slatkin, M. and Freimer, N.B. (1995), 'Microsatellite allele frequencies in humans and chimpanzees, with implications for constraints on allele size', *Mol. Biol. Evol.* Vol. 12, pp. 594–603.
53. Zhivotovsky, L.A., Feldman, M.W. and Grishechkin, S.A. (1997), 'Biased mutations and microsatellite variation', *Mol. Biol. Evol.* Vol. 14, pp. 926–933.