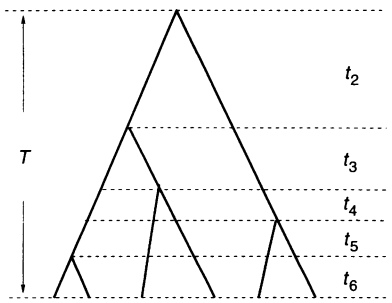

Fig. 1. An example of the genealogy of a sample of six sequences. $T = t_2 + \cdots + t_6$ is the age of the common ancestor of the sequences, and t_i is the ith coalescent time.

$$p_n(T|0) = n!\left[\prod_{i=1}^{n-1}(\theta + i)\right]\sum_{k=2}^{n}\frac{(-1)^k(\theta+2k-1)}{(k-2)!(n-k)!\prod_{i=1}^{n-1}(\theta+k+i)}e^{-k(\theta+k-1)T} \quad (4)$$

Thus, $p_n(T|0)$ depends on $\theta = 2N\mu$.

From Eq. 4, one can obtain two estimates T_{mode} and T_{mean} of T. The mode estimate T_{mode} is the value of T that maximizes the posterior probability $p_n(T|0)$, while the mean estimate T_{mean} is the expected value of T given there is no variation in the sample, that is, $T_{mean} = \int_0^\infty t \cdot p_n(t|0)dt$. In addition, the 95% confidence interval of T can be obtained from $p_n(T|0)$ as $(T_{2.5}, T_{97.5})$ where T_x is the T value such that $x\% = \int_0^T p_n(t|0)dt$. In the present situation T_{mode} is preferred over T_{mean} because the former is the most likely value of T, while the latter is more of a prediction and its computation assumes that T can be infinitely large; in reality, T must be finite. T_{95} is also of interest, because it is the 95% upper limit of T.

As the mutation rate per sequence per year has been estimated to be 0.98×10^{-6} by Dorit *et al.* (*1*), the mutation rate (μ) per sequence per generation can be estimated as $20 \times 0.98 \times 10^{-6}$ if one human generation is 20 years. However, to estimate T from Eq. 4, one needs to know the effective size N of the male human population. The data given by Dorit *et al.* do not provide enough information for a reliable estimate of N, and we therefore examine several possible values of N (Table 1).

Table 1. Estimate (1000) of age of the most recent common ancestor for male humans (T) and the 95% confidence interval for the data presented by Dorit *et al.* (*1*). Estimates are rounded to nearest thousand years.

N	T_{mode}	T_{mean}	T_{95}	Confidence interval
2.5	60.0	92.0	187.0	31.0 to 219.0
5.0	115.0	173.0	350.0	60.0 to 408.0
7.5	166.0	247.0	493.0	88.0 to 574.0
10.0	214.0	313.0	620.0	114.0 to 721.0
15.0	302.0	432.0	840.0	162.0 to 971.0
30.0	517.0	703.0	1314.0	284.0 to 1,507.0

Table 1 shows that the estimate of T and its confidence interval are dependent on N. Takahata (*4*) has suggested that the effective size of the human population (including both males and females) in the past is about 10,000. Under equal sex ratio, the effective size of the male population would be about 5,000, so that $\theta = 0.196$. Thus, T_{mode} is estimated to be 115,000 years, T_{mean} = 173,000 years, and the 95% confidence interval of T is (60,000 to 408,000 years). In addition, with 95% probability, T is smaller than 350,000 years. Our estimate T_{mean} is nearly 100,000 years less than that by Dorit *et al.* (*1*) and has a considerably smaller 95% upper limit of T. Our estimate T_{mode} is even smaller. This estimate is similar to the estimate of 143,000 years ago for the age of the MRCA of human mitochondria calculated by Horai *et al.* (*5*), though only half of that calculated by others (*6*) and is also similar to the estimates of 116,000 and 156,000 years ago that has been calculated for the age of the MRCA of humans (*7*).

Our estimate should be taken with caution because it assumes that no selective sweep on the Y chromosome has occurred in recent time. This caveat notwithstanding, it is interesting that even a DNA sample with no variation can provide much insight into human evolution.

Yun-Xin Fu
Wen-Hsiung Li
Human Genetics Center, SPH,
University of Texas,
Post Office Box 20334,
Houston, TX 77225, USA
E-mail: fu or li@hgc.sph.uth.tmc.edu

REFERENCES AND NOTES

1. R. L. Dorit, H. Akashi, W. Gilbert, *Science* **268**, 1183 (1995).
2. G. Watterson, *Theor. Popul. Biol.* **7**, 256 (1975).
3. J. F. C. Kingman, *J. Appl. Prob.* **19A**, 27 (1982); R. R. Hudson, *Theor. Popul. Biol.* **23**, 183 (1983); F. Tajima, *Genetics* **105**, 437 (1983).
4. N. Takahata, *Mol. Biol. Evol.* **10**, 2 (1993).
5. S. Horai, K. Hayasaka, R. Kondo, K. Tsugane, N. Takahata, *Proc. Natl. Acad. Sci. U.S.A.* **92**, 532 (1995).
6. R. L. Cann, M. Stoneking, A. C. Wilson, *Nature* **325**, 31 (1987); L. Vigilant, M. Stoneking, H. Harpending, K. Hawkes, A. C. Wilson, *Science* **253**, 1503 (1987).
7. M. Nei and A. K. Roychoudhury, *Evol. Biol.* **14**, 1 (1982); D. B. Goldstein, A. Ruiz Linares, L. L. Cavalli-Sforza, M. W. Feldman, *Proc. Natl. Acad. Sci. U.S.A.* **92**, 6723 (1995).

9 August 1995; accepted 19 January 1996

Dorit *et al.* (*1*) used polymorphism on the Y chromosome to infer aspects of human population history. They found an absence of sequence variation in a worldwide sample of 38 human males at a 729–base-pair intron located immediately upstream of the ZFY zinc-finger exon. They argue that, on the basis of these data, a coalescent model predicts an expected time to a most recent common ancestral male lineage of 270,000 years, with 95% confidence limits of 0 and 800,000 years.

There are errors in this report (*1*) in the application of coalescent theory. As other investigators may wish to draw inferences about the time to common ancestors, we present valid analyses from both classical and Bayesian perspectives. These lead to broadly similar point and interval estimates to those in the report (*1*). Such summary statistics do not, however, tell the full story. Likely values for the time since the common ancestor of the sampled chromosomes are substantially smaller than the point estimate of 270,000 years given in (*1*). Furthermore, the data are not particularly informative about this time—they are also consistent with much larger values than the upper estimate of 800,000 years (*1*).

Let T represent the time in years since the most recent common ancestor of the sampled sequences, N the effective population size, μ the mutation rate (per generation) of the sampled region, and D the data—the observed absence of variability. In contrast to the statement by Dorit *et al.* in (*1*), there is no simple expression for $P(D|T)$. However, given the values of N and μ, the probability $P(D)$ of the data is known (*2*)

$$P(D) = \prod_{i=1}^{37}\frac{i}{i + 2N\mu}$$

The data thus bear directly on inferences for N and μ, and only indirectly on T. For the values $\mu = 1 \times 10^{-5}$, 1.96×10^{-5} [corresponding to the value used in the report (*1*)] and 5×10^{-5}, respectively, the upper 95% confidence limits for N are 40200, 20500, and 8000.

In the coalescent model, conditional on D, the time T is $N \times G \times S$, where G is the generation time and S is the sum of 37 independent exponential random variables with respective means $2/[i(i - 1 + 2N\mu)]$, i = 2,3,. . ., 38. In particular

$$E(T|D) = NG\sum_{i=2}^{38}\frac{2}{i(i - 1 + 2N\mu)}$$

Conditioning on the data reduces the mean of T (by 20% to 40% for plausible values of N) from the value of $2NG$ used in the report (*1*). The median, mean, 5th, and 95th percentiles of the conditional distribution of T given D, for $\mu = 1.96 \times 10^{-5}$ and G = 20 years, as a function of N are shown (Fig. 1). Observe that increasing the population size *increases* values of T (*1*).

The inference concerning T in (*1*) is

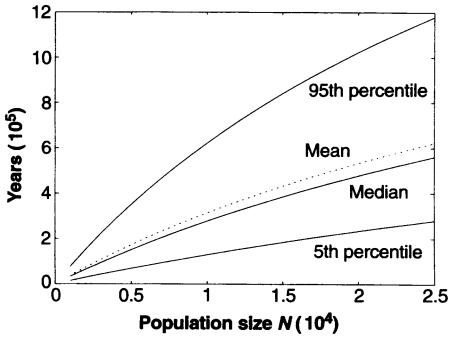

Fig. 1. Summary statistics for the conditional distribution, under the coalescent model, of the time T (in years) since the common ancestor, given a sample of 38 sequences which exhibit no variability, as a function of N, the effective population size. The generation time is assumed to be 20 years, and the mutation rate of the sequenced region per generation is taken to be 1.96×10^{-5}. Conditional distribution of T follows from equation 5.2 in (*7*).

Fig. 2. The posterior probability density function of T for various assumptions about the mutation rate μ and the effective population size N. A lognormal distribution is used to model the prior uncertainty about μ (so that $\log(\mu)$ has a normal distribution). The lognormal probability density is

$$f(x) = \frac{1}{xs\sqrt{2\pi}} \exp\left(\frac{-(\log x - m)^2}{2s^2}\right).$$

The parameters m and s were chosen to give various standard deviations, with the prior mean of μ fixed at 1.96×10^{-5}. Two different distributions were used to describe the prior information about N: (**A**) a uniform distribution and (**B**) a lognormal distribution with parameters $m = 10$ and $s = 1$. In the latter case, N has prior mode about 8,100, median 22,000 and mean 36,000. The density is at least half the modal value when N is in the interval 2,500 to 26,000. Each curve in the figure is obtained using density estimation based on 10,000 simulated values.

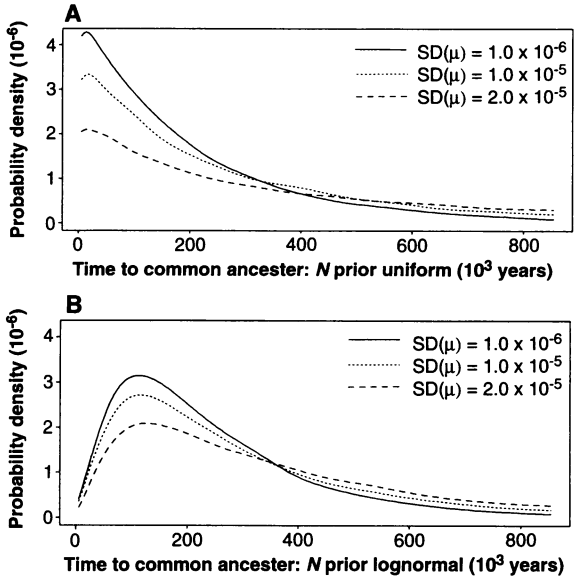

Table 1. Summary statistics of the posterior distributions illustrated in Fig. 2. SE of the means due to the finite number of simulations (10,000) are about 1% of the values. Relative simulation errors for the other statistics are broadly similar.

Prior for N	Prior SD for μ	Posterior summary statistics		
		Statistic*	T	N
Uniform	1×10^{-6}	5th	10,600	370
		median	142,000	4,800
		mean	217,000	7,300
		95th	673,000	22,600
Uniform	1×10^{-5}	5th	13,500	460
		median	199,000	6,600
		mean	347,000	11,800
		95th	1,180,000	39,000
Uniform	2×10^{-5}	5th	21,200	720
		median	391,000	13,100
		mean	890,000	30,400
		95th	3,430,000	113,000
Log-normal	1×10^{-6}	5th	49,700	1,900
		median	201,000	6,900
		mean	254,000	8,400
		95th	642,000	20,000
Log-normal	1×10^{-5}	5th	53,000	2,100
		median	234,000	7,900
		mean	324,000	10,300
		95th	891,000	26,400
Log-normal	2×10^{-5}	5th	63,400	2,400
		median	305,000	10,000
		mean	460,000	13,900
		95th	1,380,000	38,500

*5th and 95th percentiles are given.

Bayesian, with a uniform prior distribution for T. Given N, the coalescent model specifies the distribution of T, so that the uniform prior is not appropriate. Nonetheless, Bayesian inference is particularly valuable in the presence of relatively little data, and some information from other sources. The probability densities for T, conditional on the data, for various different assumptions about the pre-data uncertainty in N and μ are shown (Fig. 2). (Summary statistics of each curve in Fig. 2 are given in Table 1). If, initially, all possible values of N are regarded as equally likely (up to some large value), then a wide range of values for T is plausible. The most likely values of T after observing the data are small, around 15,000 years, a value which seems implausible in the light of our knowledge of human history. On the basis of a lognormal prior, which gives a more realistic assessment of the information available about N, the most likely, or modal, values of T are around 120,000 years. Again, a very wide range of values is plausible. The effect on inferences about T of uncertainty about the value of μ is shown (Fig. 2): The greater this uncertainty, the more plausible are large values of T. Intuitively, this is because the observed absence of variation can be explained by a smaller mutation rate, in which case the data convey less information about N and T.

In the above analyses, T is the time until the common ancestor of the *sample*. This need not be the same as "Adam," the common ancestor of all existing Y chromosomes. Under the assumptions of the coalescent model, and conditional on D, for $N\mu = 7500 \times 1.96 \times 10^{-5} \approx 0.15$ there is a probability of 0.07 that Adam will occur earlier than T (*3*). In this case, the additional time before T until Adam has mean and SD approximately NG years, which is likely to be substantial.

Under the coalescent model, N represents the "variance" effective population size, calculated as the actual number of breeding males divided by the variance of the number of male offspring of a typical male. This variance could be large if there were disparities, perhaps for reasons of social organization, in the reproductive success of different males in early human societies. If this obtained, the value of N could be substantially smaller than the actual number of breeding males in the population.

The coalescent model may be extended to allow for variation in population size and non-random mating resulting from geographical population structure. We investigated the effects of recent population expansion (*4*) for a population that was of constant size N_1 before 50,000 years ago, when it began exponential growth. For the range of parameters considered, the time to the most recent common ancestor of the sample behaves like the corresponding time for the (constant-sized) population of size N_1, plus about 42,000 years. Therefore, the model (Fig. 1) may be used to find the distribution of T. Informally, the effect of geographical structure is to increase coalescence times, often very substantially. It is thus likely that, conditional on D, non-random mating will also increase T, and the time since Adam, in contrast to the statement by Dorit *et al.* (*1*).

The analyses discussed here deal with inference for coalescence times when the data display no variability. For other data sets, for example that presented by Hammer (*5*), alternative computer-intensive methods are available (*6*).

Peter Donnelly
Departments of Statistics, and Ecology and Evolution, University of Chicago, 5734 University Avenue, Chicago, IL 60637, USA
Simon Tavaré
Departments of Mathematics and Biological Sciences, University of Southern California, Los Angeles, CA 90089-1113, USA
E-mail: stavare@gnome.usc.edu

David J. Balding
School of Mathematical Sciences, Queen Mary and Westfield College, Mile End Road, London, E1 4NS, United Kingdom
Robert C. Griffiths
Department of Mathematics, Monash University, Clayton, 3168, Australia

REFERENCES AND NOTES

1. R. L. Dorit, H. Akashi, W. Gilbert, *Science* **268**, 1183 (1995).
2. G. A. Watterson, *Theor. Popul. Biol.* **7**, 256 (1975); W. J. Ewens, *ibid.* **3**, 87 (1972).
3. Write $p(m, n)$ for the probability that a sample of size m sequences from the population has the same common ancestor as a subsample of n of the m sequences, given that the n sequences exhibit no variability. Standard arguments show that the $p(m, n)$ satisfy the recursion

$$(m(m-1)+2N\mu n)p(m, n)$$
$$=n(n-1+2N\mu)p(m-1, n-1)+$$
$$[m(m-1)-n(n-1)]p(m-1, n),$$

with initial conditions $p(m, 1) = 1$ if $m = 1$ and 0 otherwise, and $p(n, n) = 1$. We evaluated $\lim_{m\to\infty} p(m, 38)$ numerically.
4. Variable population size was modeled as follows: the population was of constant size $N_1 = \alpha N_0$ until Z years ago, when it began exponential growth to its current size N_0. The population size t years ago is $N_0\alpha^{\min(t/Z, 1)}$. We used values $Z = 50{,}000$, $N_0 = 10^8$ and 10^6, while $N_1 = 100{,}000$, 50,000, 5,000, and 1,000. We assumed $\mu = 1.96 \times 10^{-5}$. The conditional distribution of the time to the common ancestor is computed by a Monte Carlo method. In a simulation run, let $v_2, \ldots, v_{38}$ be the times while there are $2, \ldots, 38$ ancestors of the sample. These times are simulated from a coalescent model with varying population size as shown by R. C. Griffiths and S. Tavaré [*Philos. Trans. R. Soc. Lond. B* **344**, 403 (1994)]. Let $t = v_2 + \cdots + v_{38}$ be the time to the common ancestor, $w = 2v_2 + \cdots + 38v_{38}$ be the total edge length of the coalescent tree, and $q = \exp(-N_0\mu w)$ be the probability of no mutation, given the coalescent tree. The empirical distribution of the time to the most recent common ancestor from r simulation runs takes values $t_1, t_2, \ldots, t_r$ with probabilities $p_1, p_2, \ldots, p_r$ where $p_i = q_i / \sum_{j=1}^{r} q_j$, $i = 1, \ldots, r$. An estimate of $E(T|D)$ is $\sum_{j=1}^{r} t_j q_j / \sum_{j=1}^{r} q_j$.
5. M. F. Hammer, *Nature* **378**, 376 (1995).
6. R. C. Griffiths and S. Tavaré, *Stat. Sci.* **9**, 307 (1994); S. Tavaré, D. J. Balding, R. C. Griffiths, P. Donnelly, in preparation (preprint available from authors).
7. S. Tavaré, *Theor. Popul. Biol.* **26**, 119 (1984).
8. P.D. was supported in part by NSF grant DMS 95-05129 and by the Block Fund of the University of Chicago. S.T. was supported in part by NSF grants DMS 90-05833, BIR 95-04393, and NIH grant GM36232. D.J.B. was supported in part by the Science Research Fellowship scheme of the Nuffield Foundation. R.C.G. was supported in part by an Australian Research Council grant.

14 July 1995; accepted 19 January 1996

Dorit *et al.* (*1*) studied the sequence variation of an intron located in the *ZFY* gene from a sample comprising 38 sequences. Unexpectedly, the sequences did not show any variation, which means that routine methods (*2*) for analyzing such data are not applicable to this sequence.

Using coalescence theory (*3*), Dorit *et al.* argue that the MRCA of the Y chromosome existed some 270,000 years ago, with a "95% maximum estimate" of 800,000 years

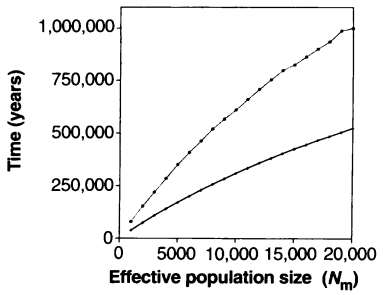

Fig. 1. Estimated times back (lower curve) to the MRCA of the Y chromosome and estimated upper 95% confidence bound (upper curve) (*7*). Abscissa represents the effective population size.

[see note 15 in (*1*)]. However, the computation is flawed. The crucial mistake (among others) is that Dorit *et al.* use an incorrect formula [see the first formula in note 15 in their report (*1*)] that does not take the effective population size of males (N_m) into account.

We have reanalyzed the data to obtain correct values (4) of the estimated times back to the MRCA for various values of N_m together with the upper 95% confidence bound (Fig. 1). If the effective population size exceeds 20,000 males, then the probability to observe no variation drops below 5% and hence it is unlikely that N_m is larger than 20,000. However, the most likely value for N_m is zero, which is unrealistic. If we assume an N_m of 5000 (*5*) then the ancestor of the Y chromosome lived approximately 170,000 years ago, with a 95% confidence interval of 0 to 350,000 years. A population size of 8500 would lead to the time estimate of 270,000 years given by Dorit *et al.* (*1*). Our estimated upper time limit (540,000 years) is considerably below their estimate of 800,000 years. Thus, we have no insights on the long-term effective population size of men. The possible range of expected times back to the father of all Y chromosomes lies between 0 and 520,000 years, if population size remains constant.

The assumption of a constant population size is extremely unrealistic for human populations. A more likely scenario is that of an exponentially growing population. Dorit *et al.* also address this question. Assuming a star phylogeny, they conclude that the MRCA existed 27,000 years ago. With the use of coalescence theory under the assumption of an exponentially growing population (*6*), we computed the expected time back to the MRCA for various growth rates, given that all sequences in the sample are identical (*7*). If the population growth rate is smaller than 0.003 per generation, then the probability of observing no variation is below 5% (Table 1).

Thus, we conclude that the growth rate of males must exceed this value. Assuming

Table 1. Estimates of expected times $E_{\theta,n}(T|X = 0)$, in years, back to the MRCA of the Y chromosome and the upper 95% confidence bound (T_{max}) for different growth rates. The analysis is based on the mutation rate given by Dorit *et al.* (*1*) and the method as outlined in note (*4*). The last column gives the probability to observe no variation in a sample of $n = 38$ sequences.

Growth rate	$E_{\theta,n}(T\|X = 0)$	T_{max}	$Pr_{\theta,n}(X = 0)$
0.001	286,000	302,000	0.0003
0.002	150,000	159,000	0.013
0.003	103,000	109,000	0.051
0.004	78,600	83,000	0.102
0.005	63,800	67,000	0.156
0.006	53,800	57,000	0.208
0.007	46,600	49,000	0.256
0.008	41,000	43,200	0.299
0.009	36,800	38,600	0.339
0.010	33,200	35,000	0.374
0.011	30,400	32,000	0.407
0.012	28,000	29,400	0.436
0.013	26,000	27,400	0.463
0.014	24,200	25,400	0.485
0.015	22,800	23,800	0.509
0.016	21,400	22,400	0.532
0.017	20,000	21,000	0.552
0.018	18,800	19,800	0.571
0.019	18,000	19,000	0.586
0.020	17,000	18,000	0.602

$r = 0.003$, we calculate the time back to the MRCA to be 103,000 years, with a 95% confidence interval of 0 to 109,000 years.

The time of 27,000 years, suggested by Dorit *et al.* (*1*) for the star phylogeny, corresponds to a growth rate of approximately $r = 0.013$. This value of r implies that roughly 32,000 years were necessary to produce N_m of today, which appears to be unrealistic (*8*).

In conclusion, coalescence theory, correctly applied, provides a plausible range of dates for the MRCA of the Y chromosome, which seems to be compatible with the current view of modern human evolution derived primarily from the analysis of mitochondrial DNA (*9*). However, to ensure a more thorough analysis of the evolution of the Y chromosome, more sequence data that also exhibit variation, are necessary. Furthermore, we have only applied two simple models about evolution of human populations. It remains to be seen how more complex scenarios of population history will affect our estimates.

Gunter Weiss
Arndt von Haeseler
Institute of Zoology, University of Munich, Post Office Box 202136, D-80021 Munich, Germany
E-mail: arndt@zi.biologie.uni-muenchen.de

REFERENCES AND NOTES

1. R. L. Dorit, H. Akashi, W. Gilbert, *Science* **268**, 1183 (1995).
2. F. Tajima, *Genetics* **123**, 585 (1989); M. Kreitman