Robert L. Dorit et al. (1) examined a world-wide sample of 38 human males and found no variation in a 729-base pair intron of the ZFY gene. Any conventional estimate of the age of the most recent common ancestor (MRCA) that is proportional to the mean number of nucleotide differences between two sequences or the number of segregating sites in the sample will give a zero value for such data, which is apparently unacceptable. To deal with this situation, Dorit et al. (1) used the Bayesian approach in conjunction with the coalescent theory of population genetics. They obtained 270,000 years ago as an estimate of the age of the most recent common ancestor, with 95% confidence limits of 0 to 800,000 years. Their approach is interesting, but the formula they derived is rough. We provide here a more rigorous method and show that the age may be only half of the estimate made by Dorit et al.

Let \( p_0(0|T) \) be the probability that a sample of \( n \) sequences contains no variation, given the age of \( T \) of their most recent common ancestor. Then the posterior probability \( p_\alpha(T) \) of \( T \), given that there is no variation in the sample, is

\[
p_\alpha(T) = \int \cdots \int \left[ \prod_{k=2}^{n} e^{-\lambda_k(k-1)} \right] \ dt_1 \cdots dt_n
\]

where \( \lambda_k \) is the prior probability of \( T \). To estimate \( T \), it is essential to obtain \( p_0(0|T) \). Wasserfenn (2) showed that the probability of no variation in a sample of size \( n \) is

\[
q_0(0) = \frac{1 - 2z \cdots (n-1)}{(1+z)(2+z) \cdots (n+z)}
\]

where \( z = 2N_q \) for a locus on \( X \) chromosome, \( N \) is the effective size of the male population, and \( \mu \) is the mutation rate per sequence per generation. Dorit et al. (1) apparently used this formula for \( p_0(0|T) \) by substituting \( T \) for \( 2N_q \), because the expected value of \( T \) is approximately equal to \( 2N \). This substitution, however, neglects the stochastic variation of \( T \) and leads to inaccurate results.

One can avoid the above problem by deriving the exact formula for \( p_0(0|T) \) using the coalescent theory (3). Let \( t_k \) be the \( k \)th coalescent time, that is, the period during which the sample has exactly \( k \) ancestral sequences (Fig. 1). The age of the MRCA of the sample is \( T = t_1 + \cdots + t_n \). According to the coalescent theory, \( t_k \) follows the exponential distribution with density \( \exp \left[ -(k-1) \right] \), where one unit of time corresponds to 2N generations. If the number of mutations in a given period is a Poisson variable, the probability that there is no mutation in a sequence during the period of \( t_k \) is \( e^{-2N\mu t_k} = e^{-2N\mu t_k} \). There are \( k \) ancestral sequences in the sample during the period of \( t_k \) (Fig. 1). Therefore, the joint probability that there is no mutation during the period of \( t_k \) and that \( t_k = t \) is

\[
e^{-2N\mu t_k} \sum_{i=0}^{\infty} e^{-2N\mu t_k} \cdot \frac{(2N\mu t_k)^i}{i!} = e^{-2N\mu t_k} \sum_{i=0}^{\infty} \frac{(2N\mu t_k)^{i+1}}{(i+1)!}
\]

The joint probability that there is no variation in the entire genealogy and that the age of the MRCA of the sample is \( T \) is given by

\[
p_0(0,T) = \int \cdots \int \left[ \prod_{k=2}^{n} e^{-\lambda_k(k-1)} \right] \ dt_1 \cdots dt_n
\]

where \( \lambda_k \) is the prior probability of \( T \). To estimate \( T \), it is essential to obtain \( p_0(0|T) \). Wasserfenn (2) showed that the probability of no variation in a sample of size \( n \) is

\[
q_0(0) = \frac{1 - 2z \cdots (n-1)}{(1+z)(2+z) \cdots (n+z)}
\]

where \( z = 2N_q \) for a locus on \( Y \) chromosome, \( N \) is the effective size of the male population, and \( \mu \) is the mutation rate per sequence per generation. Dorit et al. (1) apparently used this formula for \( p_0(0|T) \) by substituting \( T \) for \( 2N \), because the expected value of \( T \) is approximately equal to \( 2N \). This substitution, however, neglects the stochastic variation of \( T \) and leads to inaccurate results.

One can avoid the above problem by deriving the exact formula for \( p_0(0|T) \) using the coalescent theory (3). Let \( t_k \) be the \( k \)th coalescent time, that is, the period during which the sample has exactly \( k \) ancestral sequences (Fig. 1). The age of the MRCA of the sample is \( T = t_1 + \cdots + t_n \). According to the coalescent theory, \( t_k \) follows the exponential distribution with density \( \exp \left[ -(k-1) \right] \), where one unit of time corresponds to 2N generations. If the number of mutations in a given period is a Poisson variable, the probability that there is no mutation in a sequence during the period of \( t_k \) is \( e^{-2N\mu t_k} = e^{-2N\mu t_k} \). There are \( k \) ancestral sequences in the sample during the period of \( t_k \) (Fig. 1). Therefore, the joint probability that there is no mutation during the period of \( t_k \) and that \( t_k = t \) is

\[
e^{-2N\mu t_k} \sum_{i=0}^{\infty} e^{-2N\mu t_k} \cdot \frac{(2N\mu t_k)^i}{i!} = e^{-2N\mu t_k} \sum_{i=0}^{\infty} \frac{(2N\mu t_k)^{i+1}}{(i+1)!}
\]

The joint probability that there is no variation in the entire genealogy and that the age of the MRCA of the sample is \( T \) is given by

\[
p_0(0,T) = \int \cdots \int \left[ \prod_{k=2}^{n} e^{-\lambda_k(k-1)} \right] \ dt_1 \cdots dt_n
\]
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 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, \( T_{\text{mode}} \) is estimated to be 115,000 years, \( T_{\text{mean}} \) = 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 \( T_{\text{mean}} \) of nearly 100,000 years is less than that by Dorit et al. (1) and has a considerably smaller 95% upper limit of T. Our estimate \( T_{\text{mode}} \) 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 whli@hgc.sph.tmc.edu

REFERENCES AND NOTES
8. A. Cavalieri, 1989; 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 insertion 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, \( \mu \) 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 \( \mu \), the probability \( P(D) \) of the data is known (2)

\[
P(D) = \prod_{i=1}^{17} \frac{1}{i + 2N\mu}
\]

The data thus bear directly on inferences for N and \( \mu \), and only indirectly on T. For the values \( \mu = 1 \times 10^{-5}, 1.96 \times 10^{-5} \) [corresponding to the value used in the report (1)] and 5 \times 10^{-5}, respectively, the upper 95% confidence limits for N are 4200, 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 \( 2[i(i-1) + 2N\mu] \), \( i = 2, 3, \ldots, 38 \). 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, 95th, and 99th 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
Estimating the Age of the Common Ancestor of Men from the ZFY Intron


*Science* **272** (5266), 1356-1362
DOI: 10.1126/science.272.5266.1356

**References**
This article cites 11 articles, 4 of which you can access for free
http://science.sciencemag.org/content/272/5266/1356#BIBL

**Permissions**
http://www.sciencemag.org/help/reprints-and-permissions

Use of this article is subject to the Terms of Service