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

From Eq. 4, one can obtain two estimates \( T_{\text{mode}} \) and \( T_{\text{mean}} \) of \( T \). The mode estimate \( T_{\text{mode}} \) is the value of \( T \) that maximizes the posterior probability \( p_s(T|0) \), while the mean estimate \( T_{\text{mean}} \) is the expected value of \( T \) given there is no variation in the sample, that is, \( T_{\text{mean}} = \sum T \cdot p_s(T|0) \). In addition, the 95% confidence interval of \( T \) can be obtained from \( p_s(T|0) \) as \( (T_{2.5}, T_{97.5}) \) where \( T \) is the value such that \( x^{(T_{2.5})} = \int_{0}^{T} (k - 2)!(n - k) \Pi_{i=1}^{n-i} (\theta + k + i) \sum_{k=2}^{n} \frac{(-1)^{i}(\theta + 2k - 1)}{(k - 2)!} \frac{\theta^{n - k}}{n!} \frac{e^{-\theta}}{\theta^{k}} \frac{(\theta + k + 1)_{T}}{\theta^{T}} \). In the present situation \( T_{\text{mode}} \) is preferred over \( T_{\text{mean}} \), 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 \( 0.98 \times 10^{-8} \) by Dorit et al. (1), the mutation rate \( \mu \) per sequence per generation can be estimated as \( 20 \times 0.98 \times 10^{-8} \) 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 (7) and the 95% confidence interval for the data presented by Dorit et al. (1). Estimates are rounded to nearest thousand years.

<table>
<thead>
<tr>
<th>( N )</th>
<th>( T_{\text{mode}} )</th>
<th>( T_{\text{mean}} )</th>
<th>( T_{95} )</th>
<th>Confidence interval</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.5</td>
<td>60.0</td>
<td>92.0</td>
<td>187.0</td>
<td>31.0 to 219.0</td>
</tr>
<tr>
<td>5.0</td>
<td>115.0</td>
<td>173.0</td>
<td>350.0</td>
<td>80.0 to 408.0</td>
</tr>
<tr>
<td>7.5</td>
<td>166.0</td>
<td>247.0</td>
<td>493.0</td>
<td>88.0 to 574.0</td>
</tr>
<tr>
<td>10.0</td>
<td>214.0</td>
<td>313.0</td>
<td>620.0</td>
<td>114.0 to 721.0</td>
</tr>
<tr>
<td>15.0</td>
<td>302.0</td>
<td>432.0</td>
<td>840.0</td>
<td>162.0 to 971.0</td>
</tr>
<tr>
<td>30.0</td>
<td>517.0</td>
<td>703.0</td>
<td>1314.0</td>
<td>284.0 to 1507.0</td>
</tr>
</tbody>
</table>

Thus, \( \hat{T} \) is the age of the common ancestor of the sequences, and \( t \) is the \( i \)th coalescent time.

The data thus bear directly on inferences for \( N \) and \( \mu \), and only indirectly on \( T \). For the values \( \mu = 1 \times 10^{-3}, 1.96 \times 10^{-3} \) [corresponding to the value used in the report (1)] and \( 5 \times 10^{-3} \), 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 \( 2/(i - 1 + 2N\mu) \), \( i = 2,3,\ldots,38 \). In particular

\[
E(T|D) = N\sum_{i=2}^{38} 2/(i - 1 + 2N\mu)
\]

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 \( \mu = 1.96 \times 10^{-3} \) and \( G = 20 \) years, as a function of \( N \) are shown (Fig. 1). Observe that increasing the population size increases values of \( T \).

The inference concerning \( T \) in (1) is
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 \( \mu \) 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 \( \mu \) 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 \( N_1 = 7500 \times 1.96 \times 10^{-5} \approx 0.15 \) 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
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 [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 \(\tau = 0.003\), 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 \(\tau = 0.013\). This value of \(\tau\) 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.