*et al.*, *J. Mol. Biol.* **215**, 403 (1990)] and GCG software [J. Devereux, P. Haeberli, O. Smithies, *Nucleic Acids Res.* **12**, 387 (1984)]. The DNA sequences of the *ARE1* and *ARE2* genes are deposited at GenBank (P25628 and U51790, respectively).

18. KO-5′ and KO-3′ primers (GAGGGGACGAAAATTAGCCGCTATTAATTCTGGTATTGCCACCTAGACAAGAAGTAAACAGACACAGATGcaagagttcgaatctcttagc and CTATAAAGATTTAATAGCTCCACAGAACAGTTGCAGGATGCCTTAGGGTCGActacgtcgtaaggccgtttctgac, respectively; the lowercase lettering corresponds to the *LEU2* gene) were used in a polymerase chain reaction (PCR) with the *LEU2* gene as a template to produce the selectable yeast gene flanked by *ARE2* gene sequences [A. Baudin, O. Ozier-Kalogeropoulos, A. Denouel, C. Cullin, *Nucleic Acids Res.* **21**, 3329 (1993)]. This was used to transform a derivative of yeast strain 5051, heterozygous for the *are1ΔNA* allele. To identify integrants at the *ARE2* locus, we performed PCR on genomic DNA from these strains using are2-5′ (CATTGCAGTTACACGTGAATGC), are2-3′ (TAGCTCCACAGAACAGTTGCAGG), and a 3′ primer corresponding to the *LEU2* gene (L2-3′: CTCTGACAACAACGAAGTCAG).
19. P. Greenspan, E. P. Mayer, S. D. Fowler, *J. Cell Biol.* **100**, 965 (1985).
20. One to two units (at an absorbance at 600 nm) of cells were incubated in YPD or defined media containing 1 μCi/ml of [$^3$H]oleate in tyloxapol-ethanol (1:1) for 16 hours. Total lipids were prepared by hexane extraction [L. W. Parks, C. D. Bottema, R. J. Rodriguez, T. A. Lewis, *Methods Enzymol.* **111**, 333 (1985)] and analyzed by thin-layer chromatography on DC-plastikfolien kieselgel 60 plates (E-Merck, Germany). The plate was developed in hexane, diethyl ether, and acetic acid (70:30:1) and stained with iodine vapor. Incorporation of label into triglyceride and ergosterol ester was ascertained after scintillation counting and normalization to a [$^{14}$C]cholesterol internal standard and the dry weight of the cells.
21. S. L. Sturley, H. Yang, J. T. Billheimer, in preparation.
22. To overexpress the *ARE1* gene by copy number under the control of its own promoter in YEp3-16, a 2354-bp Cla I fragment from pH3(34), encompassing the entire *ARE1* gene, was blunt-ended with Klenow DNA polymerase I and introduced into the Sma I site of YEp352. To constitutively overexpress *ARE1* from the ADH promoter in pADH5-36, a 2290-bp Nar I fragment of pH3(34), starting 70 bp 5′ to the ORF, was blunt-ended with Klenow and ligated to Klenow-treated, Eco RI–digested pDC-ADH [a derivative of pS5; S. L. Sturley *et al.*, *J. Biol. Chem.* **269**, 21670 (1994)]. Increased expression of the *ARE1* transcripts, relative to that in a wild-type cell, was confirmed by Northern blot analysis.
23. C. C. Chang *et al.*, *J. Biol. Chem.* **270**, 29532 (1995).
24. G. J. Warner *et al.*, *ibid.*, p. 5772.
25. The incorporation of [1-$^{14}$C]acetate into saponified lipids was assessed as a measurement of sterol synthesis. Approximately 2 units at an absorbance of 600 nm of cells were incubated with 20 μCi of [1-$^{14}$C]acetate in 2 ml of defined media at 30°C for 3 hours and subjected to lipid saponification, hexane extraction, and thin-layer chromatography [R. Y. Hampton and J. Rine, *J. Cell Biol.* **125**, 299 (1994)]. The incorporation of counts into total sterols was assessed after scintillation counting. To normalize the estimate of sterol biosynthesis to incorporation of acetate into the fatty acid pool, we acidified the aqueous lysate remaining after hexane extraction with concentrated HCl and re-extracted it with hexane [D. Dimster-Denk, M. K. Thorsness, J. Rine, *Mol. Biol. Cell* **5**, 655 (1994)].
26. I. Tabas, D. A. Weiland, A. R. Tall, *J. Biol. Chem.* **261**, 3147 (1986).
27. M. Krieger and J. Herz, *Annu. Rev. Biochem.* **63**, 601 (1994).
28. M. E. Basson, M. Thorsness, J. Rine, *Proc. Natl. Acad. Sci. U.S.A.* **83**, 5563 (1986); S. L. Thompson, R. Burrows, R. J. Laub, S. K. Krisans, *J. Biol. Chem.* **262**, 17420 (1987).
29. We gratefully acknowledge the assistance of I. Becker, W. H. Mewes, and A. Goffeau in screening the confidential data set obtained in the European sequencing project. We thank L. A. Grivell and P. Philippsen for the provision of chromosome III DNA clones and the shotgun library of cosmid 14-21 from chromosome XIV, respectively. We thank A. Keesler and I. Tabas for helpful discussions, R. Golick for his assistance with confocal and fluorescence microscopy, N. Erdeniz for help with micromanipulations, and J. J. Rich, S. Gangloff, and A. Tinkelenberg for a critical reading of the manuscript. This work was supported in part by a Grant-in-Aid/Investigatorship from the American Heart Association (New York City affiliate) and by the Ara Parseghian Medical Research Foundation to S.L.S., NIH grants GM50237 and HG00861 to R.R., R01 AI38598 to M.B., and HL40404 to R.J.D., and by the European Community within the framework of the BIOTECH program. M.B. acknowledges support from the Johnson and Johnson Focused Giving program.

20 December 1995; accepted 11 April 1996

# TECHNICAL COMMENTS

## Estimating the Age of the Common Ancestor of Men from the *ZFY* Intron

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 $p_n(0|T)$ be the probability that a sample of $n$ sequences contains no variation, given the age $T$ of their most recent common ancestor. Then the *posterior* probability $p_n(T|0)$ of $T$, given that there is no variation in the sample, is

$$p_n(T|0)=\frac{p_n(0|T)p(T)}{\int_0^\infty p_n(0|t)p(t)dt} \quad (1)$$

where $p(T)$ is the *prior* probability of $T$. To estimate $T$, it is essential to obtain $p_n(0|T)$. Watterson (*2*) showed that the probability of no variation in a sample of size $n$ is

$$q_n(0|\theta)=\frac{1\cdot 2\cdots(n-1)}{(1+\theta)(2+\theta)\cdots(n-1+\theta)} \quad (2)$$

where $\theta$ is equal to $2N\mu$ 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 $p_n(0|T)$ 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 $p_n(0|T)$ 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 $T = t_2 + \cdots + t_n$. According to the coalescent theory, $t_k$ follows the exponential distribution with density $k(k-1)\exp[-k(k-1)t]$, 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 $e^{-\mu 2Nt_k} = e^{-\theta t_k}$. 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

$$e^{-k\theta t}k(k-1)e^{-k(k-1)t}$$

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

$$p_n(0,T) = \int\cdots\int_{t_2+\cdots+t_n=T}\left[\prod_{k=2}^{n}e^{-k\theta t_k}k(k-1)e^{-k(k-1)t_k}\right]dt_n\ldots dt_2$$

$$= n!(n-1)!\sum_{k=2}^{n}\frac{(-1)^k(\theta+2k-1)}{(k-2)!(n-k)!\prod_{i=1}^{n-1}(\theta+k+i)}e^{-k(\theta+k-1)T} \quad (3)$$

Eq. 3 is obtained by integrating with respect to coalescent times repeatedly. Because $p(0, T) = p(0|T)p(T)$, we can show that Eq. 1 becomes

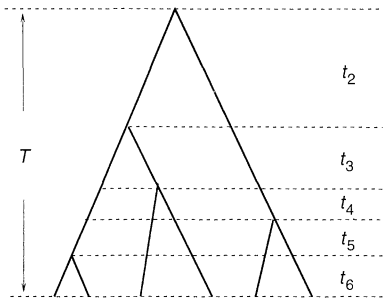


**Fig. 1.** An example of the genealogy of a sample of six sequences. $T = t_2 + \cdots + t_6$ is the age of the common ancestor of the sequences, and $t_i$ is the $i$th coalescent time.

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

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 $0.98 \times 10^{-6}$ by Dorit *et al.* (*1*), the mutation rate ($\mu$) 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, $\mu$ 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*)

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

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 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

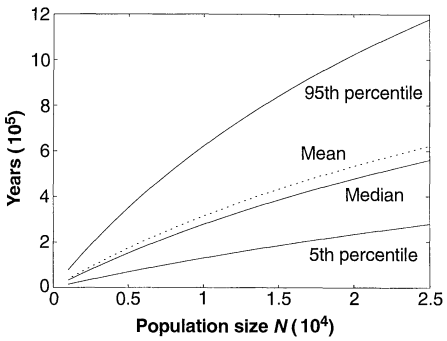


**Fig. 1.** Summary statistics for the conditional distribution, under the coalescent model, of the time $T$ (in years) since the common ancestor, given a sample of 38 sequences which exhibit no variability, as a function of $N$, the effective population size. The generation time is assumed to be 20 years, and the mutation rate of the sequenced region per generation is taken to be $1.96 \times 10^{-5}$. 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 $\mu$ and the effective population size $N$. A lognormal distribution is used to model the prior uncertainty about $\mu$ (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 $\mu$ fixed at $1.96 \times 10^{-5}$. 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.

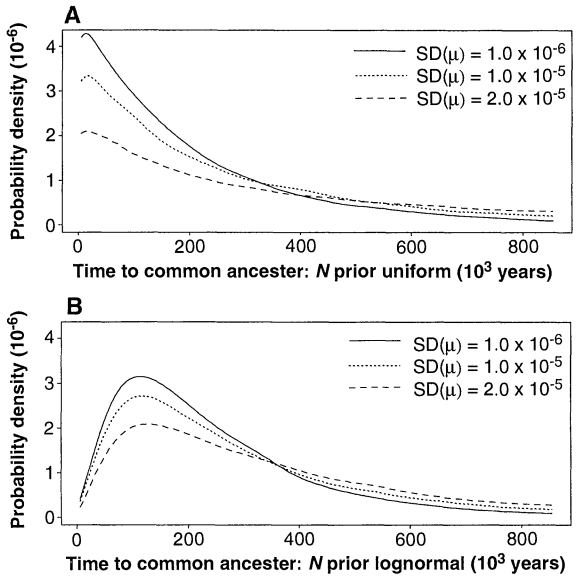


**Table 1.** Summary statistics of the posterior distributions illustrated in Fig. 2. SE of the means due to the finite number of simulations (10,000) are about 1% of the values. Relative simulation errors for the other statistics are broadly similar.

| Prior for N | Prior SD for μ | Posterior summary statistics | | |
|---|---|---|---|---|
| | | Statistic* | T | N |
| Uniform | $1 \times 10^{-6}$ | 5th | 10,600 | 370 |
| | | median | 142,000 | 4,800 |
| | | mean | 217,000 | 7,300 |
| | | 95th | 673,000 | 22,600 |
| Uniform | $1 \times 10^{-5}$ | 5th | 13,500 | 460 |
| | | median | 199,000 | 6,600 |
| | | mean | 347,000 | 11,800 |
| | | 95th | 1,180,000 | 39,000 |
| Uniform | $2 \times 10^{-5}$ | 5th | 21,200 | 720 |
| | | median | 391,000 | 13,100 |
| | | mean | 890,000 | 30,400 |
| | | 95th | 3,430,000 | 113,000 |
| Log-normal | $1 \times 10^{-6}$ | 5th | 49,700 | 1,900 |
| | | median | 201,000 | 6,900 |
| | | mean | 254,000 | 8,400 |
| | | 95th | 642,000 | 20,000 |
| Log-normal | $1 \times 10^{-5}$ | 5th | 53,000 | 2,100 |
| | | median | 234,000 | 7,900 |
| | | mean | 324,000 | 10,300 |
| | | 95th | 891,000 | 26,400 |
| Log-normal | $2 \times 10^{-5}$ | 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 $\mu$ 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 $N\mu = 7500 \times 1.96 \times 10^{-5} \approx 0.15$ 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

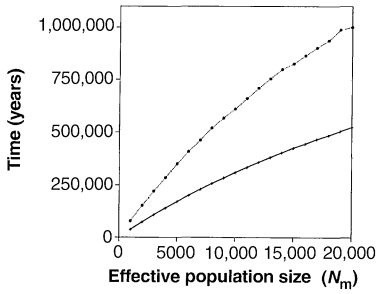


**Fig. 1.** Estimated times back (lower curve) to the MRCA of the Y chromosome and estimated upper 95% confidence bound (upper curve) (*7*). Abscissa represents the effective population size.

[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

and R. Hudson, *ibid.* **127**, 565 (1991); J. Felsenstein, *Genet. Res. Camb.* **59**, 139 (1992). Y. X. Fu, *Genetics* **136**, 685 (1994). S. T. Sherry *et al.*, *Hum. Biol.* **66**, 761 (1994).
3. R. R. Hudson, *Oxf. Surv. Evol. Biol.* **7**, 1 (1990).
4. Let $N_m$ and $\mu$ be the effective population size and the mutation rate, respectively. $X$ denotes the number of variable sites. The probability to observe no variation in a random sample of size $n$ drawn from a Wright-Fisher population is given by

$$\Pr_{\theta,n}(X = 0) = \frac{\Gamma(n)\cdot\Gamma(1 + \theta)}{\Gamma(n + \theta)} \quad (1)$$

where $\theta = 2N_m\mu$ [S. Tavaré, *Theor. Popul. Biol.* **26**, 119 (1984)]. The probability of the time $T = t$ back to the MRCA, conditional on seeing no variation in the sample equals according to Bayes theorem

$$\Pr_{\theta,n}(T = t|X = 0) = \frac{\Pr_{\theta,n}(T = t, X = 0)}{\Pr_{\theta,n}(X = 0)} \quad (2)$$

where $\Pr_{\theta,n}(T = t, X = 0)$ is the joint probability of time $T$ back to the MRCA and $X = 0$. This joint probability was estimated by running 1,000,000 Monte-Carlo simulation for each value of $\theta$. On the basis of estimated values of $\Pr_{\theta,n}(T = t, X = 0)$, we can infer

$$E_{\theta,n}(T|X = 0) = \int_0^\infty t\cdot\Pr_{\theta,n}(T = t|X = 0)dt \quad (3)$$

the expected time until the sample coalesces to a single sequence, given that $X = 0$. We similarly estimated the upper limit of the 95% confidence region. The formulae given above depend only on the compound parameter $\theta$. Dorit *et al.* (*1*) estimated a substitution rate of 0.135% per million years for the intron. Assuming this value is approximately correct, we have used the corresponding substitution rate of $\mu = 9.8 \cdot 10^{-7}$ per sequence and per year.
5. A. R. Rogers and L. B. Jorde, *Hum. Biol.* **67**, 1 (1995).
6. M. Slatkin and R. R. Hudson, *Genetics* **129**, 555 (1991).
7. Using theory (*6*), we have generated 1,000,000 Monte Carlo genealogies for each choice of growth rate $r$. We further assume that the effective population size of males today is about 1,000,000,000. For each $r$ we computed the same quantities as defined in (*4*). Now, the important parameter is not $N_m$ but rather $r$. Unfortunately, we are not aware of a closed formula to compute $\Pr_{\theta,n}(X = 0)$ in this situation, hence this quantity was estimated by simulations.
8. Computation is based on a current effective population size of $1 \cdot 10^9$ men. If the effective size is smaller, then the estimate of the time back to the MRCA will only slightly decrease (data not shown).
9. S. Pääbo, *Science* **268**, 1141 (1995).
10. Supported by a grant from DFG to A. v. H. Discussion with S. Pääbo is also greatly acknowledged.

29 June 1995; accepted 19 January 1996

Dorit *et al.* (*1*) compare a 729–base pair intronic sequence of the Y-linked *ZFY* gene from one orangutan, one gorilla, one chimpanzee (*Pan paniscus*, Genbank accession no. U24117), and 38 humans. On the basis of this comparison, they constructed a phylogenetic tree representing the evolution of the *ZFY* locus. The maximum parsimony tree obtained indicates that chimpanzee-bonobo *ZFY* is more closely related to human *ZFY* than to the same locus from gorillas. This gene tree matches the topology of gene trees developed for mitochondrial DNA and other nuclear DNA sequences (*2*). However, the species-level phylogeny of these taxa remains controversial, as studies of other loci have obtained discordant results (*3*).

The phylogenetic conclusions presented by Dorit *et al.* lead to an important dilemma. If their evolutionary history for the Y chromosome is correct, and if it accurately reflects the evolutionary history of the genera *Homo*, *Pan*, and *Gorilla*, then significant aspects of the generally accepted model of human evolution must be incorrect. It is generally agreed that humans, chimpanzees, and gorillas form a closely related group of species, though the details of the relationships within this clade have been difficult to resolve (*3*, *4*). It is even more broadly agreed that modern humans are more closely related to the extinct genus *Australopithecus* and its more derived relative *Paranthropus* than to either chimpanzees or gorillas (*5*). The two genera *Australopithecus* and *Paranthropus* are represented by hundreds of fossils from Pliocene and early Pleistocene geological formations in eastern and southern Africa (*5*). We consider the idea that humans share a last common ancestor with *Australopithecus* more recently than with *Pan* or *Gorilla* to be firmly established. The *ZFY* gene tree suggests that humans, chimpanzees, and gorillas share a last common ancestor about 5 million years ago (Ma), and that the divergence of the human lineage from the chimpanzee lineage occurred approximately midway between that date and the present. The gene tree has three nucleotide substitutions occurring along the internode that represents the common ancestor of *Homo* and *Pan*, and has an average of 2.5 nucleotide substitutions in the two lineages resulting from the *Homo-Pan* split. With the use of the rate of *ZFY* evolution that Dorit *et al.* observed, we calculate that the *ZFY* data suggest that the *Homo-Pan* divergence occurred 2.54 Ma.

However, extensive and widely accepted paleontological and geological research has shown that the genus *Australopithecus* was present in Africa and was using fully bipedal locomotion more than 3 Ma (*5*, *6*). Recent finds from Ethiopia have extended the range of *Australopithecus* or other closely related taxa back to between 4.0 and 4.4 Ma (*7*). Thus, the model of Dorit *et al.* dates the human-chimpanzee divergence subsequent to the origin of *Australopithecus*.

We see three ways out of this dilemma. First, one could postulate that humans are more closely related to *Pan* than to *Australopithecus*. This hypothesis requires that we accept one of the following conclusions: (i) that chimpanzees evolved their current knuckle-walking locomotion and other primitive morphological features from an ancestor exhibiting many derived features of bipedalism as well as other cranial and post-cranial characters found in more derived human ancestors, or (ii) that a long list of derived cranial and post-cranial characters believed to be homologous in human ancestors and in *Australopithecus* actually arose independently through convergent evolution.

Second, one could propose that the topology of Dorit's *ZFY* gene tree is correct, but that the absolute dates are wrong. If the divergence of *Gorilla* ancestors from the common *Homo-Pan* ancestor was about 8 to 9 Ma rather than 5 Ma, the *Homo-Pan* divergence would fall at about 4 to 5 million years, and possibly resolve the problem. But this model has two important implications: (i) It suggests that the divergence of orangutans from the other three hominoids must have been considerably earlier than the date used by Dorit *et al.* (14 Ma). The date would probably be significantly earlier than 20 Ma, and this is unlikely given other evidence (*8*). (ii) This solution implies that the rate of evolution of the *ZFY* sequence is lower than Dorit *et al.* calculated, and therefore pushes the date of their inferred human Y chromosome coalescence substantially earlier in time. The new conclusion, that this coalescence occurred roughly 0.4 Ma, with 95% confidence limits of 0 to over 1 million years, would dramatically reduce the impact of the *ZFY* data in relation to the question of modern human origins.

The third solution to the dilemma is to accept the second most parsimonious tree for the *ZFY* gene sequence. Dorit *et al.* report that the most parsimonious tree, which links *Homo* and *Pan* to the exclusion of *Gorilla*, requires 70 mutational steps. They also state that a tree 72 mutational steps in length is the next most parsimonious, and that this tree reconstructs the evolution of the *ZFY* locus (and therefore the Y chromosome) as a trichotomous divergence. The date of this trichotomy would be about 5 Ma, roughly coincident with the earliest known fossils attributable to *Australopithecus* or closely related taxa. Given the alternatives, we favor this third solution to the dilemma and suggest that the *ZFY* locus provides an interesting illustration of two general principles: (i) that parsimony is a useful and indeed indispensible heuristic tool for evolutionary biologists, but that it should not be assumed that all DNA sequence evolution necessarily occurred in the most parsimonious manner, particularly when other reconstructions that require only a small number of additional mutational events are available, and (ii) that the most complete understanding of evolutionary history results from the careful integration of all relevant information. The fields of molecular systematics and paleontology are each important to the study of human

phylogeny, and our models of human evolution gain in depth and strength when we consider all the evidence together. Scenarios that are most parsimonious for the one dataset are not necessarily the most parsimonious when viewed globally.

*Jeffrey Rogers*
*Paul B. Samollow*
*Anthony G. Comuzzie*
*Department of Genetics, Southwest Foundation for Biomedical Research, Post Office Box 760549, San Antonio, TX 78245, USA*
*E-mail: jrogers@darwin.sfbr.org*

## REFERENCES

1. R. L. Dorit, H. Akashi, W. Gilbert, *Science* **268**, 1183 (195).
2. S. Horai, K. Hayasaka, R. Kondo, K. Tsugane, N. Takahata, *Proc. Natl. Acad. Sci. U.S.A.* **92**, 523 (1995); W. Bailey *et al.*, *Mol. Phylog. Evol.* **1**, 97 (1992).
3. J. Rogers, *J. Hum. Evol.* **25**, 201 (1989); K. J. Livak, J. Rogers, J. B. Lichter, *Proc. Natl. Acad. Sci. U.S.A.* **92**, 427 (1995); W. Bailey, *Evol. Anthropol.* **2**, 100 (1993).
4. C. Groves, in *Comparative Primate Biology, Vol. 1: Systematics, Evolution and Anatomy*, D. R. Swindler and J. Erwin, Eds. (A. R. Liss, New York, 1986), pp. 187–217.
5. W. K. Kimbel, T. D. White, D. C. Johanson, *Am. J. Phys. Anthropol.* **64**, 337 (1984); B. Wood, *Nature* **355**, 783 (1992); H. M. McHenry, *Proc. Natl. Acad. Sci. U.S.A.* **91**, 6780 (1994).
6. M. Leakey and R. L. Hay, *Nature* **278**, 317 (1979); H. M. McHenry, *J. Hum. Evol.* **15**, 177 (1986).
7. T. D. White. G. Suwa, B. Asfaw, *Nature* **371**, 306 (1994); G. WoldeGabriel *et al.*, *ibid.*, p. 330.
8. J. Kappelman *et al.*, *J. Hum. Evol.* **21**, 61 (1991).

27 June 1995; accepted 19 January 1996

*Response*: In our report (*1*), we derived a possible age for the last ancestor of the sampled Y chromosomes by using a variety of statistical approaches. We found (i) no variation in a 729-base-long intron on the Y chromosome in a worldwide sample of 38 males and (ii) a mutation rate of 0.135% per Ma, estimated from sequencing this intron in chimpanzee, gorilla, and orangutan. We then used theory-dependent arguments to place our results in the context of current debates about the history of the Y chromosome and the evolution of modern *Homo sapiens*.

In principle, the lack of variation we observe reflects the common ancestry of this region of the Y chromosome. The mutation rate suggests that one should be able to estimate how many changes would be expected for any given time elapsed since the MRCA, or, conversely, what the expectation of the time to the MRCA should be, given no observed changes. However, the MRCA calculation is entirely model driven, and we discuss two simple, but extreme, models to put a range on the expectations.

The simplest model is the "star" phylogeny, where each of the 38 individuals is seen to represent a separate line of descent from the MRCA. This model approximates a scenario where the species spreads completely (and quickly) around the world immediately after the MRCA. It is also a good approximation for a picture of rapid exponential growth of the human population. Under this model, the expected time to common ancestry is 27,000 years, with a 95% confidence limit for a deepest time of 80,000 years. Although the "star" model makes certain extreme assumptions, such as the simultaneous and rapid colonization of the entire world, the very short times predicted by the "star" phylogeny show that if the lines of descent are separated, the observed mutation-fixation rate requires a very recent common ancestor, compatible with the most recent spread of *H. sapiens* around the world 40–60,000 years ago. It is a model of this type which should be compared to the multiregional hypothesis, which postulates that the relevant spread around the world occurred 1 to 2 million years ago, and hence that the lines of descent have been separated since then.

Our second model, the coalescent phylogeny, assumes a small, equilibrium effective population throughout all (or almost all) of human history. On the basis of the size of this equilibrium population, the probability of common ancestry can be estimated by coalescing the lineages one by one to a deepest bifurcation. (The many short final lineages in this model are a consequence of the assumption of a fixed $N_e$). Under this model of lineage bifurcation, the time to the last common ancestor of the sampled Y chromosomes is likely to be larger than under the assumptions of the star phylogeny. We used a Bayesian interpretation of this model to estimate an expected time of 270,000 years, and commented in note 15 of our report that we estimated $N_e$ to be 7500 males by linking the expected value of $T$ and $N_e$.

This model, with its built-in assumption of equilibrium effective populations and the small $N_e$ that is required under this scenario by the data, is also probably an unrealistic description of the entire course of human evolution, which involves a gradual spread around the world and an increasing population. The point of presenting (*1*) these two models was to provide a range of estimates for the real time to a last common male ancestor which is likely to bracket the correct value. In our view, there was not enough experimental data presented in our report to justify an extensive discussion of intermediate models, although we note that subsequent papers (*2*) based on variational data have arrived at intermediate estimates of coalescence time.

Fu and Li present a clear discussion of one of the issues surrounding our estimates. They correctly point out that we have used an approximation to the total coalescent time, estimated as the sum of the expectations of the individual coalescence times. In practice, that approximates their integral, 3, by an integral over each of the $t_i$'s independently. This gives an estimate for $P_n(0,T)$ that is larger than the correct one, and also ties together $T$ and $N$. Nevertheless, our data do permit an estimate of the effective population size (using the Watterson formula, listed in Fu and Li as equation 2). Just as in our report, a Bayesian argument will estimate $P(N|0)$ from $P(0|N)/P(0)$ if all $N$'s are equiprobable *a priori*. Knowing the mutation rate, and assuming a generation time of 20 years, the $N_{exp}$ is 6750, with an upper bound $N_{95\%}$ of 20,000. When these values, estimated directly from our data, are then used to estimate $T$—the time to the MRCA—we derive a $T_{mean}$ of approximately 90% of the value we originally report. As Fu and Li point out, their exact handling of the data still produces time estimates for the most part smaller than ours. Because we were using the coalescence argument as a crude approximation to an oldest time, we are gratified by their comments.

The comments by Donnelly *et al.* and Weiss and von Haeseler explore the consequences on the coalescence model of varying assumptions about the mutation rate, the effective population size, or the dynamics of population growth. Not unexpectedly, the model is sensitive to such parameters. Thus, for example, the coalescent model presented by Weiss and von Haeseler incorporates exponential population growth and yields estimates of the time to the MRCA intermediate between the star phylogeny and equilibrium effective population size scenarios. Similarly, the incorporation of an underlying sampling variance in the mutation rate, or the use of mutation rates other than the one we empirically derive (as presented by Donnelly *et al.*), will necessarily increase the uncertainty in any estimate of the age of the MRCA. These authors also comment on the fact that, under a coalescence model, an increase in the assumed effective population size results in an increase in the time to coalescence, as would be expected given the relationship between $N$ and $T$ in the model. In real terms, however, an increased actual population size makes the probability of finding no polymorphism in our sample less and less likely, unless the time to the MRCA is pushed closer and closer to the present.

Although developments in coalescent

models will allow the incorporation of more complex sampling and population dynamic scenarios, the data presented in our report did not justify such additional considerations. Similarly, while more sequence data, from this and other loci, will be required before the full evolutionary history of Y chromosomes and of our species can be deciphered, our report was both an attempt to initiate this evolutionary reconstruction and an example of how the absence of variation represents an evolutionary signal in its own right.

Finally, we wish to clarify a point raised by Rogers *et al.* Although the most parsimonious tree that can be derived from our data does in fact place the chimpanzee-human split after the branching off of the gorilla lineage (supported by two characters), we were careful to state, in note 10 of the report, that the next shortest tree describes an unresolved trichotomy. When we calculated an expected mutation rate for this intron (note 11), we assumed such a trichotomy, and used independent estimates of branching times of 5MY for the chimpanzee-human, gorilla-human, and chimpanzee-gorilla splits (14MY for the splitting off of orangutan). We then averaged over all possible pairwise comparisons to obtain a mean mutation rate.

Given the small number of changes taking place along the branches and nodes of this gene tree, our data should not be used in a molecular clock form to estimate the age of the interspecific splits, as was done by Rogers *et al.* If one considers only the numbers of changes, the observed numbers (5, 10, and 11) for the human-chimpanzee, human-gorilla, and chimpanzee-gorilla comparisons, respectively, are not significantly different from the 8, 8, and 8 expected from a trichotomy ($\chi^2 = 2.75$).

**Robert L. Dorit**
*Department of Biology,*
*Yale University,*
*165 Prospect Street,*
*New Haven, CT 06511, USA*
**Hiroshi Akashi**
*Department of Ecology and Evolution,*
*University of Chicago,*
*Chicago, IL 60637, USA*
**Walter Gilbert**
*Department of Molecular and Cellular Biology,*
*Harvard University,*
*Cambridge, MA 02138, USA*

## REFERENCES

1. R. L. Dorit, H. Akashi, W. Gilbert, *Science* **268**, 1183 (1995).
2. M. F. Hammer, *Nature* **378**, 376 (1995); L. S. Whitfield, J. E. Sulston, P. N. Goodfellow, *ibid.*, p. 379; S. A. Tishkoff *et al.*, *Science* **271**, 1380 (1996).

26 January 1996; revised 27 March 1996; accepted 2 April 1996

# Correlates of Protective Viruses Damaging to HIV Infection

Barton F. Haynes *et al.* (*1*) state correctly that concentrations of human immunodeficiency virus (HIV) are low and of cytotoxic T lymphocytes (CTLs) are high in people who are "nonprogressors." Therefore, they argue, our proposals—that HIV is essentially not a lytic virus and that immunosuppression may be caused by virus-specific $CD8^+$ T cell-mediated immunopathology that destroys infected antigen-presenting and T cells—do not apply. This is an incorrect conclusion drawn from our views, because the example of the nonprogressor with a low HIV load and high CTL response does fit into our balance-scheme between the two extremely rare cases that Haynes *et al.* quote from our proposal (*2*). If efficient CTL killing (plus neutralising antibody) eliminates HIV completely before it can be integrated into many cells, HIV negativity and immunity will result. If high CTL activity (plus antibody) controls infection early and efficiently, long-term nonprogression will result (with potential incubation times of more than 30 years). If the balance is in the middle, the average of 8 to 10 years necessary for development of the disease will result; if the growth of HIV is less, but still somewhat controlled, immunopathology will develop quicker to cause disease. The other extremely unbalanced state occurs when no T cell responses are available, or T cells become exhausted by too wide an infection, which probably is enhanced by the developing immunopression. This latter extreme situation would correspond to a "healthy" hepatitis B virus carrier state. The dynamic balance between virus and immunopathology depends on the discussed various host (human lymphocyte antigen, interferon, and so forth) and virus (escape mutants, susceptibility to interferon, and so forth) parameters; their combination differs from patient to patient, yielding the wide spectrum of disease patterns and disease kinetics. The view that disease is caused by immunopathology—that is, by the damaging effects of the protective immune response—has important implications for therapy and prevnetion. Accordingly, enhancement of an immune response that is beneficial when the HIV load is low, may be damaging and enhance disease when virus has already spread widely. Absence of evidence that HIV is directly lytic in vivo must encourage us to search for evidence, or absence, of an important role of immunopathology in AIDS pathogenesis.

**Rolf M. Zinkernagel**
**Hans Hengartner**
*Institute of Experimental Immunology,*
*University Hospital of Zurich,*
*Schmelzbergstrasse 12,*
*CH-8091 Zurich, Switzerland*
*Email: expimm.@usz.unizh.ch*

## REFERENCES

1. B. F. Haynes, G. Pantaleo, A. S. Fauci, *Science* **271**, 324 (1996).
2. B. Odermatt, M. Eppler, T. P. Leist, H. Hengartner, R. M. Zinkernagel, *Proc. Natl. Acad. Sci. U.S.A.* **88**, 8252 (1991); R. M. Zinkernagel and H. Hengartner, *Immunol. Today* **15**, 262 (1994).

7 February 1996; accepted 21 March 1996

*Response*: The remark in our report (*1*) about CTLs was not meant to imply that the elegant and provocative hypothesis of Zinkernagel and Hengartner (*2*) was invalid. Rather, it was intended to point out that it is difficult to hypothesize that CTLs are either immunopathogenic or protective only on the basis of quantitative differences in the CTL response. For example, if one examines CTL responses in HIV-infected individuals in early stages of the disease, it is not unusual to observe high frequencies of HIV-specific cytotoxicity despite the fact that the vast majority of these individuals will ultimately progress in their disease. Quantitation of the CTL response early in the course of HIV disease does not seem to predict progression of disease. In contrast, qualitative differences in the CTL response as reflected by recognition of variable versus conserved epitopes, and the mobilization of a broader (as opposed to a more restricted) CTL repertoire, may determine whether a CTL response will be pathogenic or protective.

**Barton F. Haynes**
*Department of Medicine,*
*Duke University Medical Center,*
*Durham, NC 27710, USA*
**Giuseppe Pantaleo**
**Anthony S. Fauci**
*National Institute of Allergy and Infectious Disease,*
*National Institutes of Health,*
*Bethesda, MD 20892, USA*

## REFERENCES

1. B. F. Haynes, G. Pantaleo, A. S. Fauci, *Science* **271**, 324 (1996).
2. R. M. Zinkernagel and H. Hengartner, *Immunol. Today* **15**, 262 (1994).

15 March 1996; accepted 22 March 1996