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

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

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

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

where θ is equal to $2N\mu$ for a locus on Y chromosome, N is the effective size of the male population, and μ 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 kth 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

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

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

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

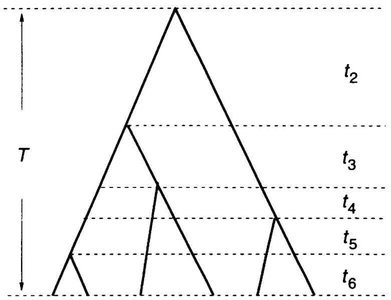

Fig. 1. An example of the genealogy of a sample of six sequences. $T = t_2 + \cdots + t_6$ is the age of the common ancestor of the sequences, and t_i is the *i*th coalescent time.

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

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, $T_{\text{mean}} = \int_0^\infty t \cdot p_n(t|0)dt$. 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 $2/[i(i - 1 + 2N\mu)]$, $i = 2,3,\ldots, 38$. 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, 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

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

Y.-X. Fu, W.-H. Li, P. Donnelly, S. Tavaré, D. J. Balding, R. C. Griffiths, G. Weiss, A. von Haeseler, J. Rogers, P. B. Samollow, A. G. Comuzzie, R. L. Dorit, H. Akashi and W. Gilbert

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