Alu Evolution in Human Populations: Using the Coalescent to Estimate Effective Population Size

Stephen T. Sherry,*,† Henry C. Harpending,*,‡ Mark A. Batzer† and Mark Stoneking*

** Department of Anthropology, The Pennsylvania State University, University Park, Pennsylvania 16802, † Department of Pathology, Department of Biometry and Genetics, Neuroscience Center of Excellence, Stanley S. Scott Cancer Center, Louisiana State University Medical Center, New Orleans, Louisiana 70112 and ‡ Department of Anthropology, University of Utah, Salt Lake City, Utah 84112*

Manuscript received May 1, 1997
Accepted for publication September 3, 1997

ABSTRACT

There are estimated to be ~1000 members of the Ya5 *Alu* subfamily of retroposons in humans. This subfamily has a distribution restricted to humans, with a few copies in gorillas and chimpanzees. Fifty-seven Ya5 elements were previously cloned from a HeLa-derived randomly sheared total genomic library, sequenced, and screened for polymorphism in a panel of 120 unrelated humans. Forty-four of the 57 cloned *Alu* repeats were monomorphic in the sample and 13 *Alu* repeats were dimorphic for insertion presence/absence. The observed distribution of sample frequencies of the 13 dimorphic elements is consistent with the theoretical expectation for elements ascertained in a single diploid cell line. Coalescence theory is used to compute expected total pedigree branch lengths for monomorphic and dimorphic elements, leading to an estimate of human effective population size of ~18,000 during the last one to two million years.

ALU insertions are primate-specific genetic elements that mobilize via the process of retroposition (reviewed in DEININGER and BATZER 1993, 1995). Approximately 500,000 *Alu* sequences exist within the human genome, representing ~5% of the genome by mass. *Alu* elements are thought to be ancestrally derived from 7SL RNA, and thus they do not encode the reverse transcriptase necessary for element duplication. Rather, they are thought to exploit the reverse transcriptase encoded by L1 elements (FENG *et al.* 1996; MORAN *et al.* 1996). *Alu* elements are generally believed to be noncoding because of their neutral substitution rate, small size (300 bp) and lack of 5′ *cis*-acting RNA polymerase III transcription signals (DEININGER and BATZER 1993). The exceptions to the latter are the very few transcriptionally active "master" or source genes that are currently generating new family members (DEININGER *et al.* 1992).

This ongoing amplification process has produced families of elements shared at all primate taxonomic hierarchies. Subfamilies are defined by mutations that accumulate in their respective transcriptionally active copies (SLAGEL *et al.* 1987; BRITTEN *et al.* 1988; JURKA and SMITH 1988; SHEN *et al.* 1991). In humans the process has produced families of approximately 500 to 2000 elements (BATZER *et al.* 1995), now designated as "young" with subfamilies Ya5, Ya8 and Yb8 denoting groups of elements derived from separate related transcriptionally active "master" or source genes (BATZER *et al.* 1996). The young *Alu* elements are randomly distributed throughout the genome with a slight positive bias for AT-rich regions (ARCOT *et al.* 1995a, 1996).

At least 13 members of Ya5 subfamily are so recent in origin that they are polymorphic in the human gene pool (ARCOT *et al.* 1995a, and references therein). Six of these variable elements are found at relatively high frequency in populations ($P > 0.66$), two at moderate frequency ($0.33 < P < 0.66$) and five are found at low frequency ($P < 0.33$). Here, we derive the expected distribution of these frequencies for the special case of ascertainment in a single diploid genome. We then show how the ratio of polymorphic to fixed elements provides an estimate of the effective size of our species. This estimate does not require knowledge of the mutation or insertion rate.

MATERIALS AND METHODS

Previous ascertainment of Ya5 subfamily members: Members of the Ya5/8 subfamilies were isolated from a HeLa (ATCC CCL2) randomly sheared total genomic library constructed in bacteriophage λZAPII (Stratagene) as previously described (BATZER and DEININGER 1991). Fifty-seven individual clones were previously sequenced using internal *Alu* specific primers to obtain the sequence of the 5′ and 3′ flanking nucleotide sequences (BATZER *et al.* 1991; ARCOT *et al.* 1995b,c, 1996, 1997). Based on the sequence information, oligonucleotide primers specific for the 5′ and 3′ flanking unique sequences were designed and used to direct PCR amplification and sequencing of the 57 individual *Alu* inserts. Genotypes were determined for each element by inspecting PCR reaction products for the presence of a 100-bp fragment (the pre-integration site) and/or a 400-bp fragment (the pre-integration site and the *Alu* element). The genotypes of the HeLa cell line and a panel of 122 human individuals (U.S.

Corresponding author: Stephen Sherry, Department of Biometry and Genetics, Louisiana State University Medical Center, 1901 Perdido St., New Orleans, LA 70112 E-mail: ssherr@lsumc.edu

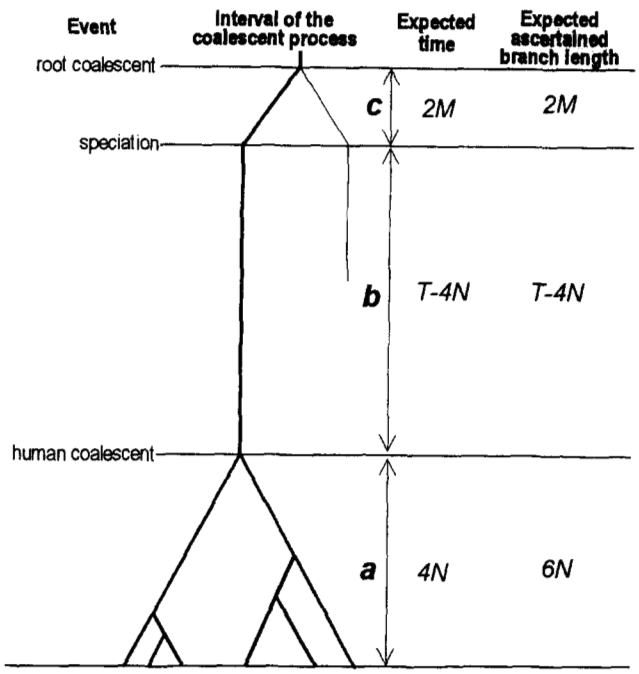

FIGURE 1.—Intervals in which *Alu* elements may accumulate during the coalescent process and their relationship to effective population size. Interval (*a*) estimates the time to the human coalescent ($4N$) by noting that the total number of recent insertions (observed dimorphic elements) is $6N\lambda$ in expectation in an ascertained genealogy; (*b*) estimates the time from speciation to the human coalescent by the number of monomorphic elements; (*c*) estimates the time forward from the root coalescent to a speciation event as $2M$ in the ancient past. See main text for procedures to calculate the expected duration and total branch length of each interval.

Caucasians, African Americans, and Hispanic Americans) were determined by fractionation of PCR products on a 2% agarose gel (BATZER *et al.* 1991; PERNA *et al.* 1992). The chromosomal location of each *Alu* element was determined by PCR amplification of human-rodent monochromosomal somatic cell hybrid DNAs (BATZER *et al.* 1991; ARCOT *et al.* 1995b,c, 1996, 1997). Thirteen of the 57 *Alu* elements were determined to be dimorphic in the human DNA samples; the Genbank/EMBL accession numbers for these elements are listed in Table 1. The Genbank/EMBL accession numbers for the 44 monomorphic loci are as follows: U18387–U18400, X54175–X54177, X54179–X54180, U67208–U67210, U67212–U67220, and U67222–U67235.

Modeling element accumulation in the genome: To model the accumulation of *Alu* insertions in the nuclear genome we assume that there is an effectively infinite number of potential insertion sites. The insertion rate (λ) is assumed to be constant in time, and individual elements are never completely lost once an insertion has occurred. Under this model individuals who share a specific *Alu* insertion inherit the element from a common ancestor. Since both the direct repeats and unique sequences that flank the element are specific to target insertion sites, this assumption is confirmed experimentally by examining every element's flanking region sequence for homology across individuals and heterogeneity across elements (ARCOT *et al.* 1995c; KNIGHT *et al.* 1996).

Figure 1 is a schematic tree of the history of a nuclear gene in a sample of humans: the actual details of any tree will vary from locus to locus. Fixed elements that are absent from pongid genomes have accumulated in the interval between the human coalescence and the ancestral hominid/pongid coalescence (Figure 1, intervals *b* and *c*). Dimorphic elements must be more recent insertions, having occurred in the bottom interval of the tree (Figure 1, interval *a*). We can compute the expected total length for both these portions of the tree, measuring the length of the branches in units of generations. To compute the total length of interval *a*, we note that a tree of n chromosomes has $n-1$ epochs. Each epoch j has an expected duration of $4N/j(j-1)$ generations (FELSENSTEIN 1992). Since there are j lineages in existence at epoch j, the expected contribution to the total tree length from epoch j is $4N/(j-1)$, and the total contribution from all epochs (sum over all j) is $4N\Sigma_{j=1}^{n-1}(1/j)$, a familiar expression due to WATTERSON (1975). Dimorphic elements, then, accumulate in a genealogy with total expected length

$$4N \sum_{i=1}^{n-1} \frac{1}{i} \qquad (1)$$

in a stationary population where N is the effective size during the nuclear coalescent process. In our sample with $n \approx 240$ (described above) this would correspond to an expected value of approximately $24N$ generations, since the sum in Watterson's formula is ~5.9. Monomorphic elements, in contrast, accumulate on the single lineage that connects interval *a* (Figure 1) to the deeper pongid coalescent as noted above.

The expected distribution of element frequencies in a sample: The distribution of alleles by frequency is referred to as the allele frequency spectrum (EWENS 1972; KIMURA and OHTA 1975; HUDSON 1990). Here we consider two spectra: the distribution of all elements currently segregating in the population and, second, the distribution of elements represented in a diploid genome, such as in a library used for screening.

Both frequency spectra may be derived from coalescent theory. We consider frequency to be the probability that a lineage (with the element) at some epoch j has k descendants in the sample, so that k/n is the frequency of the element in the sample. This probability, as pointed out by FELSENSTEIN (1992), is easily derived from Polya's urn model. At any epoch, the probability distribution of the number of descendant lineages in a sample, from a single lineage, is

$$\frac{(j-1)(n-k-1)!(n-j)!}{(n-1)!(n-j-k+1)!}$$

(FELLER 1957). At the top of the tree ($j = 2$) the probability distribution is uniform over the interval $k = 1 \ldots (n-1)$, while at $j = n$ all of the probability is at $k = 1$. Summing over epochs, the relative frequency of insertions with population frequency p is proportional to $1/p$ (NEI 1987), and thus the population frequency spectrum, $\Phi(p)$, is proportional to $1/p$.

The screened library frequency spectrum, however, is different. Many elements currently segregating in the population at low frequency are expected to be absent from the screened library. Thus, the screening process creates *ascertainment bias* in the distribution of frequencies of sampled elements, positive toward higher frequencies. We write the screened-library frequency spectrum $\Phi_x(p)$ for the distribution of frequencies to indicate that our sample of elements was ascertained in x number of haploid genomes. $\Phi_x(p)$ is simply the population distribution, $\Phi(p)$, conditioned on the probability of an element being observed in x haploid genomes. Since our data were ascertained in a genomic library constructed from a single diploid cell line, we will consider the distribution of the $\Phi_2(p)$ spectrum next.

FU (1995) showed that the expected number of inserted elements to be found in exactly i haploid genomes from a sample of n haploid genomes is $4N\lambda/i$. For each element of

TABLE 1

Dimorphic *Alu* element frequencies and chromosomal location

Locus	U.S. caucasian frequency	African American frequency	Hispanic American frequency	Average frequency	Chromosome number	GenBank/EMBL accession number
HS2.43	0.12	0.03	0.05	0.07	*1*	U57005
HS3.23	0.87	0.99	0.73	0.87	*7*	U57004
HS4.32	0.63	0.51	0.46	0.53	*12*	U57006
HS4.59	0.74	0.67	0.69	0.70	*9*	X55925
HS4.65	0.01	0.16	0.09	0.09	*9*	U57007
HS4.69	0.28	0.30	0.20	0.26	*6*	U57008
HS4.75	0.66	0.82	1.00	0.83	*3*	U57009
HSA.25	0.20	0.38	0.05	0.22	*8*	X55929
HSB.65	0.68	0.72	0.48	0.57	*11*	U18396
HSC2N4	1.00	0.98	1.00	0.99	*9*	X54181
HSC4N4	1.00	0.98	0.97	0.98	*11*	X54178
HS2.25	0.23	0.13	0.18	0.20	*13*	U67211
HS4.14	0.86	0.70	0.85	0.83	*1*	U67221

type i (that is, with i copies in a sample of n), the probability of observation in at least one of two haploid genomes is

$$\frac{i}{n}\frac{i-1}{n-1} + \frac{i}{n}\left(1 - \frac{i-1}{n-1}\right) + \frac{i-1}{n-1}\left(1 - \frac{i}{n}\right)$$

$$= \frac{n(2i-1) - i^2}{n(n-1)} \approx \frac{i}{n}\left(2 - \frac{i}{n}\right).$$

Therefore, the expected number of type i elements is

$$\frac{1}{i}(4N\lambda)\frac{i}{n}\left(2 - \frac{i}{n}\right),$$

and summing over all types of elements yields the expected number of dimorphic *Alu* elements in our screened library as

$$4N\lambda \sum_i \frac{1}{i}\frac{i}{n}\left(2 - \frac{i}{n}\right) \approx 6N\lambda.$$

This is approximately one-fourth the total expected tree length if ascertainment were complete, *i.e.*, carried out in each individual of the sample. In other words since we only identify insertions that were present in one genome (a *HeLa* cell line), we have identified roughly one-fourth of the total number of dimorphic *Alu* elements that we would observe if we carried out the ascertainment in all 122 individuals.

Estimating recent human effective population size: Under our assumption that *Alu* insertions have occurred at a constant rate within the human genome, we can estimate expected values for the total branch length in intervals *a*, *b*, and *c* of Figure 1. Insertions that occurred in intervals *b* and *c* are fixed in our sample. Interval *b* is simply the time from the human nuclear coalescent back to the speciation event in the common chimp/human ancestral species. Interval *c* is the coalescent time in the ancestral species, which is in expectation twice the effective size M of that population, time being measured in generations. We can then use published estimates of the time since speciation in the common chimp-human ancestor and of the effective size of that population to estimate the human nuclear effective size.

We assume a generation time of 20 years throughout. This is less than estimates of the current human generation time but is probably reasonable for most of the Pliocene and Pleistocene. We also assume that the chimp-human speciation event occurred 4.5 million years ago and that the effective size of the ancestral species was 100,000 individuals (TAKAHATA *et al.* 1995). These assumptions imply that $a + b + c$ (Figure 1) is 8.5 million years or 425,000 generations. Since the expected duration of interval a is $4N$ generations, the duration of interval $b + c$ is $(425{,}000 - 4N)$ generations, and this is proportional to the number of fixed *Alu* insertions. The number of *Alu* insertions polymorphic in our sample should be proportional to $6N$. Using a moment estimator, we set these expected values equal to the observed number of sites of each category to obtain our estimate of N, the human nuclear effective size.

RESULTS

Previous ascertainment of Ya5 family members: The 57 selected clones of independent elements were genotyped in both the HeLa cell line and the screening panel of 122 human individuals. Forty-four elements were monomorphic in the sample and 13 were dimorphic. Nucleotide sequences and chromosomal locations for all 57 elements were reported previously (BATZER *et al.* 1991; ARCOT *et al.* 1995b,c, 1996, 1997). All clones were further screened in a panel of 15 non-human primates and determined to be restricted to the human genome (BATZER *et al.* 1994, 1996; ARCOT *et al.* 1995a–c). Locus designations, allele frequencies, chromosomal locations and Genbank/EMBL accession numbers are reported for these elements in Table 1.

Stationary population history with ascertainment: To test the hypothesis of constant population size in the recent past, the observed dimorphic element frequencies were tabulated into three intervals (Table 2) for a chi-square test of goodness of fit. The expected number of sites in the first interval, that is for $P < 1/3$, was computed as

$$\frac{13\sum_{i=1}^{41}\left(2 - \frac{i}{122}\right)}{\sum_{i=1}^{122}\left(2 - \frac{i}{122}\right)},$$

TABLE 2

Chi-square test of goodness of fit of a dimorphic element distribution under a hypothesis of constant population size in the most recent $4N$ generations of the coalescent

Element frequency	Observed	Expected	Contribution to χ^2
0.67–0.99	6	3.370	2.052
0.34–0.66	2	4.333	1.256
0.01–0.33	5	5.296	0.016
χ^2 value			3.325[a]

[a] NS, $P = 0.20$, d.f. = 2.

and similarly for the other two intervals. The chi-square value of 3.325 was nonsignificant ($Pr = 0.20$, d.f. = 2). These data do not reject a model of stationary history in the ancient past. The late Pleistocene expansion described in ROGERS and HARPENDING (1992) and HARPENDING *et al.* (1993) is too recent to be visible with this system.

Estimating human effective population size: Our assumptions imply that $(425{,}000 - 4N)\lambda = 44$ and $6N\lambda = 13$, where λ is the insertion rate for our ascertained fraction of all *Alu* insertions. Solving these yields our estimate of $N = 17{,}500$. If the ancestral chimp/human population size M was close to zero, our estimate of N would become 9000. Our estimate is slightly greater than the conventional figure of 10,000 for the nuclear effective size of humans, perhaps because the ascertainment mechanism means that we are weighting the top of the coalescent tree very heavily with the estimation procedure. Nearly half the 13 polymorphic elements we found occurred (in expectation) during the topmost interval of the coalescent when there were only two lineages.

Our results suggest that low ($\sim 10^4$) estimates of human effective size are not simply artifacts of an upper Pleistocene bottleneck but reflect our descent from a subpopulation of archaic humans that was small for most of the middle and upper Pleistocene. While the standard error of our estimate of N is on the order of 30% or more of its value (the standard error of the ratio 13:44), the implication of our data is that it was on the order of 10,000 as we look further backward toward the human nuclear coalescent.

DISCUSSION

Prior analysis of *Alu* element diversity has described the degree of population structure for these elements in various human groups (BATZER *et al.* 1994, 1996, and references therein; ARCOT *et al.* 1995a), or modeled the dynamics of element family evolution with respect to population structure and element insertion rate (TACHIDA and IIZUKA 1993). This analysis differs in its use of the conditional spectrum to estimate effective population sizes in the Middle Pleistocene. The spectrum properties outlined in this report derive from the fact that the (−) state is ancestral to (+), which permits us to time-order insertion events by state. Without this property the ancestral state is uncertain and the distribution must be "folded" across the frequency of 0.5 to reflect this uncertainty. Examples of this latter case would include most traditional nucleotide sequence applications of the infinite site model.

The estimation of recent human effective size requires that the source Ya5 *Alu* element is older than the hominid/pongid coalescent event, that all of the monomorphic elements are restricted to the human lineage, and finally that the source gene was fixed in the ancestral species. The first requirement has been satisfied by prior work (LEEFLANG *et al.* 1992, 1993a,b) that demonstrated that five additional Ya5 elements have been found in gorillas and chimpanzees. One of the five is restricted to gorillas, three are restricted to chimpanzees, and one is orthologous in humans, chimpanzees and gorillas. The second requirement is met by testing orthologous positions of all 57 Ya5 elements in chimpanzees and other non-human primates by PCR. All 57 *Alu* repeats show a single copy of the preintegration site, indicating that the site is empty in chimpanzees and other non-human primates (BATZER *et al.* 1994, 1996; ARCOT *et al.* 1995a–c). If the source element was polymorphic in human ancestors during the last one to two million years, then the effective insertion rate was less in intervals *b* and *c*, which would bias our estimate of human effective size upward. Therefore the small effective size given by our method cannot be a consequence of a polymorphic master gene during hominid evolution. Our assumption of insertion rate constancy is difficult to verify experimentally. While the retroposition process was known to be more active in early primate evolution (as determined by copy number per family) (SHEN *et al.* 1991; SARROWA *et al.* 1997), the hierarchical structure of mutations within subfamilies is consistent with this assumption (DEININGER and BATZER 1993).

In contrast to our estimate of N, TAKAHATA (1993) concluded that a long-term effective size of 100,000 was consistent with data on HLA variation at the major histocompatibility complex. This figure, however, refers to the later Cenozoic and there is no disagreement with our estimate for the Pleistocene portion of our species' history. Later AYALA (1995) suggested that the effective size of humans during the Pleistocene must have been on the order of 10^5 rather than 10^4, but subsequent comparative analysis of the pattern of divergence in HLA exons and introns has concluded that the HLA data may be explained with a recent effective size of 10,000 and diversifying selection on only 10 ancestral allelic lineages (ERLICH *et al.* 1996), thus placing our estimates of N, in agreement with this and prior nuclear DNA studies of recent human effective size.

The figures reported here are larger than the upper

end of previous mitochondrial DNA (mtDNA)-based estimates of effective size. Estimates of 500–3000 females (ROGERS and HARPENDING 1992; HARPENDING *et al.* 1993; SHERRY *et al.* 1994) should be one-half of the sizes considered here due to the mode of mtDNA inheritance. With these effective sizes and the generation time assumptions used earlier, the mtDNA figure describes population size ~200,000 years ago while our estimate describes population size at one-half to two million years ago. The disagreement between the two figures suggests a mild hourglass contraction of human effective size during the last interglacial since 6000 is very different from 18,000. On the other hand our results also deny the hypothesis that there was a severe hourglass contraction in the number of our ancestors in the late middle and upper Pleistocene. If humans were descended from some small group of survivors of a catastrophic loss of population, then the distribution of ascertained *Alu* polymorphisms would show a preponderance of high frequency insertions (unpublished simulation results). Instead the suggestion is that our ancestors were not part of a world network of gene flow among archaic human populations but were instead effectively a separate species with effective size of 10,000–20,000 throughout the Pleistocene.

Considering the question of the hominid/pongid divergence, the ratio of segregating to fixed sites can give us a rate-free estimate of the time to the human coalescent as a fraction of the total time to the root hominid/pongid coalescent. We estimate this ratio to be one-sixth. This estimate is in good agreement with the estimate of one sixth reported in an independent study of nuclear restriction fragment length polymorphism data (MOUNTAIN and CAVALLI-SFORZA 1994). We note in passing that because elements segregate independently our ratio estimate should converge to the true population value as more elements are identified and characterized. In time the size of this ratio could become one of our stronger inferences about hominid evolution.

The diploid ascertainment method used here, which initially identified and characterized novel *Alu* elements in the human genome, is efficient for finding common elements. Thus rather than the action of selection as some have proposed (BRITTEN 1994), the natural explanation of the high levels of heterozygosity observed for these elements is to be found in the screening method. Considering the polymorphic fraction of our sample at 0.22, and an estimated total copy number of 2000 members for the Ya5, Ya8 and Yb8 subfamilies (BATZER *et al.* 1995), a pool on the order of 400 variable elements remains to contribute to the $\Phi_2(p)$ distribution. If we could ascertain further elements in a larger sample of individuals, we could also examine the more recent population dynamics of humans including the population growth during the last interglacial suggested by the mitochondrial reconstructions of human population history. Other problems for study include the need for replicate studies in other primate species to establish comparative data for African non-human primate effective sizes.

We thank M. NEI, A. CLARK, K. WEISS, A. ROGERS, A. WALKER, J. KURLAND, R. SHERWOOD and two anonymous reviewers for comments on an earlier version of this manuscript. This research was supported in part by a Hill Fellowship from the Penn State Department of Anthropology and grants from the National Science Foundation (SBR-9318826) and the Wenner-Gren Foundation for Anthropological Research (5761) to S.T.S.

LITERATURE CITED

ARCOT, S. S., T. H. SHAIKH, J. KIM, L. BENNETT, M. ALEGRIA-HARTMAN *et al.*, 1995a Sequence diversity and chromosomal distribution of "young" Alu repeats. Gene **163:** 273–278.

ARCOT, S. S., J. J. FONTIUS, P. L. DEININGER and M. A. BATZER, 1995b Identification and analysis of a 'young' polymorphic Alu element. Biochim. Biophys. Acta **1263:** 99–102.

ARCOT, S. S., Z. WANG, J. L. WEBER, P. L. DEININGER and M. A. BATZER, 1995c Alu repeats: a source for the genesis of primate microsatellites. Genomics **29:** 136–144.

ARCOT, S. S., A. W. ADAMSON, L. E. LAMERDIN, B. KANAGY, P. L. DEININGER *et al.*, 1996 Alu fossil relics—distribution and insertion polymorphism. Genome Res. **6:** 1084–1092.

ARCOT, S. S., M. M. DEANGELIS, S. T. SHERRY, A. W. ADAMSON, J. E. LAMERDIN *et al.*, 1997 Identification and characterization of two polymorphic Ya5 Alu repeats. Mutat. Res. Genomics **382:** 5–11.

AYALA, F. J., 1995 The myth of Eve: molecular biology and human origins. Science **270:** 1930–1936.

BATZER, M. A., and P. L. DEININGER, 1991 A human-specific subfamily of Alu sequences. Genomics **9:** 481–487.

BATZER, M. A., V. A. GUDI, J. C. MENA, D. W. FOLTZ, R. J. HERRERA *et al.*, 1991 Amplification dynamics of human-specific Alu family members. Nucleic Acids Res. **19:** 3619–3623.

BATZER, M. A., M. STONEKING, M. ALEGRIA-HARTMAN, H. BAZAN, D. H. KASS *et al.*, 1994 African origin of human-specific polymorphic Alu insertions. Proc. Natl. Acad. Sci. USA **91:** 12288–12292.

BATZER, M. A., C. M. RUBIN, U. HELLMANN-BLUMBERG, M. ALEGRIA-HARTMAN, E. P. LEEFLANG *et al.*, 1995 Dispersion and insertion polymorphism in two small subfamilies of recently amplified human Alu repeats. J. Mol. Biol. **247:** 418–427.

BATZER, M. A., P. L. DEININGER, U. HELLMANN-BLUMBERG, J. JURKA, D. LABUDA *et al.*, 1996 Standardized nomenclature for Alu repeats. J. Mol. Evol. **42:** 3–6.

BRITTEN, R. J., 1994 Evolutionary selection against change in many Alu repeat sequences interspersed through primate genomes. Proc. Natl. Acad. Sci. USA **91:** 5992–5996.

BRITTEN, R. J., W. F. BARON, D. B. STOUT and E. H. DAVIDSON, 1988 Sources and evolution of human Alu repeated sequences. Proc. Natl. Acad. Sci. USA **85:** 4770–4774.

DEININGER, P. L., and M. A. BATZER, 1993 Evolution of retroposons, pp. 157–196 in *Evolutionary Biology*, Vol. 27, edited by M. K. HECHT, R. J. MACINTYRE, and M. T. CLEGG. Plenum Press, New York.

DEININGER, P. L., and M. A. BATZER, 1995 SINE master genes and population biology, pp. 43–60 in *The Impact of Short Interspersed Elements (SINEs) on the Host Genome*, edited by R. J. MARAIA. R. G. Landes Company, Austin, TX.

DEININGER, P. L., M. A. BATZER, C. A. HUTCHINSON III and M. H. EDGELL, 1992 Master genes in mammalian repetitive DNA amplification. Trends Genet. **8:** 307–311.

ERLICH, H. A., T. F. BERGSTROM, M. STONEKING and U. GYLLENSTEN, 1996 HLA sequence polymorphism and the origin of humans. Science **274:** 1552–1554.

EWENS, W. J., 1972 The sampling theory of selectively neutral alleles. Theor. Popul. Biol. **3:** 87–112.

FELLER, W., 1957 *An Introduction to Probability Theory and its Applications*, Ed. 2. John Wiley and Sons, New York.

FELSENSTEIN, J., 1992 Estimating effective population size from samples of sequences: inefficiency of pairwise and segregating sites as compared to phylogenetic estimates. Genet. Res. **59:** 139–147.

FENG, Q., J. V. MORAN, H. H. KAZAZIAN JR. and J. D. BOEKE, 1996

Human L1 retrotransposon encodes a conserved endonuclease required for retrotransposition. Cell **87:** 905–916.

FU, Y. X., 1995 Statistical properties of segregating sites. Theoret. Popul. Biol. **48:** 172–197.

HARPENDING, H. C., S. T. SHERRY, A. R. ROGERS and M. STONEKING, 1993 The genetic structure of ancient human populations. Curr. Anthropol. **34:** 483–496.

HUDSON, R. R., 1990 Gene genealogies and the coalescent process, pp. 1–44 in *Oxford Surveys in Evolutionary Biology,* Vol. 7, edited by D. J. FUTUYMA and J. ANTONOVICS. Oxford University Press, Oxford.

JURKA, J., and T. SMITH, 1988 A fundamental division in the Alu family of repeated sequences. Proc. Natl. Acad. Sci. USA **85:** 4775–4778.

KIMURA, M., and T. OHTA, 1975 Distribution of allelic frequencies in a finite population under stepwise production of neutral alleles. Proc. Natl. Acad. Sci. USA **72:** 2761–2764.

KNIGHT, A., M. A. BATZER, M. STONEKING, H. K. TIWARI, W. D. SCHEER *et al.,* 1996 DNA sequences of Alu elements indicate a recent replacement of the human autosomal genetic complement. Proc. Natl. Acad. Sci. USA **93:** 4360–4364.

LEEFLANG, E. P., W.-M. LIU, C. HASHIMOTO, P. V. CHOUDARY and C. W. SCHMID, 1992 Phylogenetic evidence for multiple Alu source genes. J. Mol. Evol. **35:** 7–16.

LEEFLANG, E. P., I. N. CHESNOKOV, and C. W. SCHMID, 1993a Mobility of short interspersed repeats within the chimpanzee lineage. J. Mol. Evol. **37:** 566–572.

LEEFLANG, E. P., W.-M. LIU, I. N. CHESNOKOV, and C. W. SCHMID, 1993b Phylogenetic isolation of a human Alu founder gene: drift to new subfamily identity [corrected]. J. Mol. Evol. **37:** 559–565.

MORAN, J. V., S. E. HOLMES, T. P. NAAS, R. J. DEBERARDINIS, J. D. BOEKE *et al.,* 1996 High frequency retrotransposition in cultured mammalian cells. Cell **87:** 917–927.

MOUNTAIN, J. L. and L. L. CAVALLI-SFORZA, 1994 Inference of human evolution through cladistic analysis of nuclear DNA restriction polymorphisms. Proc. Natl. Acad. Sci. USA **91:** 6515–6519.

NEI, M., 1987 *Molecular Evolutionary Genetics.* Columbia University Press, New York.

PERNA, N. T., M. A. BATZER, P. L. DEININGER and M. STONEKING, 1992 Alu insertion polymorphism: a new type of marker for human population studies. Human Biol. **64:** 641–648.

ROGERS, A. R., and H. HARPENDING, 1992 Population growth makes waves in the distribution of pairwise genetic differences. Mol. Biol. Evol. **9:** 552–569.

SARROWA, J., D. CHANG and R. MARAIA, 1997 The decline in human *Alu* retroposition was accompanied by an asymmetric decrease in SRP9/14 binding to dimeric *Alu* RNA and increased expression of small cytoplasmic *Alu* RNA. Mol. Cell. Biol. **17:** 1144–1151.

SHEN, M. R., M. A. BATZER and P. L. DEININGER, 1991 Evolution of the master Alu gene(s). J. Mol. Evol. **33:** 311–320.

SHERRY, S. T., A. R. ROGERS, H. HARPENDING, H. SOODYALL, T. JENKINS *et al.,* 1994 Mismatch distributions of mtDNA reveal recent human population expansions. Hum. Biol. **66:** 761–775.

SLAGEL, V., E. FLEMINGTON, V. TRAINA-DORGE, H. BRADSHAW and P. DEININGER, 1987 Clustering and subfamily relationships of the Alu family in the human genome. Mol. Biol. Evol. **4:** 19–29.

TACHIDA, H., and M. IIZUKA, 1993 A population genetic study of the evolution of SINEs. I. Polymorphism with regard to the presence or absence of an element. Genetics **133:** 1023–1030.

TAKAHATA, N., 1993 Allelic genealogy and human evolution. Mol. Biol. Evol. **10:** 2–22.

TAKAHATA, N., Y. SATTA and J. KLEIN, 1995 Divergence time and population size in the lineage leading to modern humans. Theor. Popul. Biol. **48:** 198–221.

WATTERSON, G. A., 1975 On the number of segregating sites in genetical models without recombination. Theor. Popul. Biol. **7:** 256–276.

Communicating editor: W-H. LI