J Mol Evol (2004) 59:478–487
DOI: 10.1007/s00239-004-2639-2

JOURNAL OF MOLECULAR EVOLUTION

Ancestral Population Sizes and Species Divergence Times in the Primate Lineage on the Basis of Intron and BAC End Sequences

Yoko Satta,[1] Michael Hickerson,[1,2] Hidemi Watanabe,[3] Colm O'hUigin,[4] Jan Klein[5]

[1] Department of Biosystems Science, Graduate University for Advanced Studies, Hayama, Kanagawa 240-0193, Japan
[2] Department of Biology, Duke University, Durham, NC 90338, USA
[3] Department of Bioinformatics and Genomics, Nara Institute of Science and Technology, Ikoma, Nara 630-0101, Japan
[4] National Cancer Institute, National Institutes of Health, Bethesda, MD, USA
[5] Abteilung Immungenetik, Max-Planck-Institut fuer Biologie, Correnstrasse 42, D-72076, Tuebingen, Germany

Received: 5 October 2003 / Accepted: 14 April 2004 [*Reviewing Editor:* Dr. Magnus Nordborg]

Abstract. The effective sizes of ancestral populations and species divergence times of six primate species (humans, chimpanzees, gorillas, orangutans, and representatives of Old World monkeys and New World monkeys) are estimated by applying the two-species maximum likelihood (ML) method to intron sequences of 20 different loci. Examination of rate heterogeneity of nucleotide substitutions and intragenic recombination identifies five outrageous loci (*ODC1*, *GHR*, *HBE*, *INS*, and *HBG*). The estimated ancestral polymorphism ranges from 0.21 to 0.96% at major divergences in primate evolution. One exceptionally low polymorphism occurs when African and Asian apes diverged. However, taking into consideration the possible short generation times in primate ancestors, it is concluded that the ancestral population size in the primate lineage was no smaller than that of extant humans. Furthermore, under the assumption of 6 million years (myr) divergence between humans and chimpanzees, the divergence time of humans from gorillas, orangutans, Old World monkeys, and New World monkeys is estimated as 7.2, 18, 34, and 65 myr ago, respectively, which are generally older than traditional estimates. Beside the intron sequences, three other data sets of orthologous sequences are used between the human and the chimpanzee comparison. The ML application to these data sets including 58,156 random BAC end sequences (BES) shows that the nucleotide substitution rate is as low as $0.6–0.8 \times 10^{-9}$ per site per year and the extent of ancestral polymorphism is 0.33–0.51%. With such a low substitution rate and short generation time, the relatively high extent of polymorphism suggests a fairly large effective population size in the ancestral lineage common to humans and chimpanzees.

Key words: Ancestral population size — Nucleotide substitution rate heterogeneity — Primate phylogeny — Species divergence time

Introduction

Molecular genetic techniques, combined with rigorous statistical methods based on population genetic models that incorporate inherent stochasticity of nucleotide substitution processes or coalescence processes of genes in a population (Kingman 1982; Tajima 1983), allow us to answer questions regarding species divergence times and extents of ancestral polymorphism. Since orthologous genes from two species must have diverged before the divergence of the species, the divergence time of genes always exceeds that of the species. Hence, if we use the gene divergence as a representative of the species diver-

Correspondence to: Yoko Satta; *email:* satta@soken.ac.jp

gence and calculate the nucleotide substitution rate, we inevitably overestimate the rate. It is therefore important to determine to what extent the gene divergence time exceeds the species divergence time. The key factor is the effective size of the ancestral species, because the excess is determined by the coalescence process of ancestral lineages of orthologous genes. Several methods have been developed to estimate the ancestral population size together with the species divergence time (for review, see Edwards and Beerli 2000; Takahata and Satta 2002). These include the trichotomy method (Nei 1987; Wu 1991), the two-species maximum likelihood (ML) method (Takahata et al. 1995), and more generalized ML methods (Yang 1997, 2002; Rannala and Yang 2003; Wall 2003).

The trichotomy method uses genealogies of orthologous genes in three closely related species, such as humans, chimpanzees, and gorillas. Depending on the coalescence process of genes in the ancestral population of the two most closely related species, the genealogies may not be identical to the species phylogeny. The mean coalescence time [$E(t)$] is $2N$ generations for a pair of genes in the ancestral population with effective size N. If N is large compared to the time interval (T) between the two successive species divergences, the coalescence is likely to have taken place before the three species descended from the common ancestor. If this happens, the gene genealogy can differ from the species phylogeny in two-thirds of the cases so that the incompatibility probability is given by $(2/3)e^{-T/(2N)}$ (Nei 1987). When the incompatibility probability and T are at hand, we can estimate N. In practice, the incompatibility probability must be estimated by comparisons of gene genealogies at a large number of different loci. However, sampling errors as well as recombination and/or multiple hits of nucleotide substitutions may obscure a true gene genealogy. In addition, it is difficult to determine the time interval T accurately. These uncertainties make the N estimate by the trichotomy method suggestive at best.

The two-species ML method uses pairs of orthologous loci sampled from two species. The method divides the nucleotide divergences into two categories: the nucleotide divergences that have occurred before and after the speciation. If the time elapsed since speciation is t_s years and the coalescence time in the ancestral population is t generations, the divergence time of a pair of orthologous loci is $t_s + tg$ years, where the generation time in the ancestral species is g years. If the nucleotide substitution rate per site per year is μ, the number (k_i) of substitutions at the ith locus accumulated in the $t_s + tg$ interval is $(t_s + tg)\ \mu L_i$, where L_i stands for the number of nucleotides compared. For a large number of orthologous locus pairs, $t_s\mu$ is constant over all pairs but $tg\mu$ differs from locus to locus according to the exponential distribution with both mean and standard deviation $2Ng\mu$ (Takahata et al. 1995; Takahata and Satta 1997). Based on this principle, the most likely estimates of t_s and t can be obtained to fit the variation of k_i among different loci. In order for the method to yield accurate estimates, however, two conditions must be fulfilled: the μ remains constant among different loci and the sites in each pair of orthologous loci are not shuffled by intragenic recombination. In practice, unfortunately, these conditions are often not fulfilled. The heterogeneity of μ across loci enlarges the variance of k_i, leading to an overestimation of the ancestral polymorphism [$x = 2E(t)\ g\mu = 4Ng\mu$] and an underestimation of the species divergence time ($y = 2t_s\mu$). On the other hand, recombination reduces the variance of k_i and underestimates the ancestral polymorphism (Yang 1997, 2002; Takahata and Satta 2002; Wall 2003).

To satisfy the assumed constant rate of nucleotide substitutions among different loci, synonymous substitutions may be suitable (Kimura 1983) and be used in the two-species ML method (Takahata and Satta 1997; Takahata 2001). However, it turns out that the use of synonymous substitutions has raised several problems. The nucleotide diversity may vary among loci because of linkage to selected sites (Hartl and Clark 1997, pp. 184–185) or biased mutation pressure (Bielawski et al. 2000 and references therein). In addition, the number of synonymous sites at a locus is generally small and tends to be underestimated by frequently used methods if nucleotide substitutions are biased toward transitions (Nei and Kumar 2000, p. 57). To avoid these problems, we may use intron or intergenic sequences (Chen and Li 2001).

In the present study, we use intron sequences of 20 loci and make two-species ML estimates of the ancestral population size and species divergence time for pairs of six primate species, i.e., humans, chimpanzees, gorillas, orangutans, and representatives of Old World monkeys (OWMs) and New World monkeys (NWMs). For the human–chimpanzee pair, we also apply the two species ML method to exon sequences of 37 loci (Takahata 2001), 53 intergenic sequences (Chen and Li 2001), and a set of 58,156 human–chimpanzee pairs of BAC End Sequences (BES; Fujiyama et al. 2002).

Materials and Methods

Intron Sequences

The intron data set was taken from O'hUigin et al. (2002; Table 1), which contains noncoding sequences such as the 5′ or 3′ untranslated regions, promoter regions, and introns. Since functional constraints against transcription or translation regulation may operate on parts of the nonintron regions, we used only introns. In the case of genes containing more than one intron, we concatenated these intron sequences. There are 20 loci for which intron sequences are available in six primate species (or taxa): humans, chimpanzees, gorillas, orangutans, macaques or baboons, and tamarins or marmosets. Since OWMs and NWMs are monophyletic to each other and to other primates, sequences from macaques/baboons and tamarins/marmosets were used as a representative of OWMs and NWMs, respectively. For simplicity, humans, chimpanzees, gorillas, orangutans, OWMs, and NWMs are abbreviated H, C, G, O, M, and T, respectively, throughout the text. Although another large set of 53 hominoid intergenic sequences is also available (Chen and Li 2001), we did not use it for two reasons. First, the set lacks OWM and NWM sequences. The divergence time of OWMs and NWMs is still controversial (Pilbeam 1984; Martin 1993; Kumar and Hedges 1998; Goodman et al. 1998; Takahata 2001; Glazko and Nei 2003; Hasegawa et al. 2003) and its estimation is one of the aims of the present study. Second, the sequences we used are about 800 bp long on average. They are longer than in Chen and Li (2001) (~500 bp on average) and therefore less prone to stochastic errors.

Sequence Alignment and Phylogenetic Analysis

Sequences were aligned by the Clustal W program (Thompson et al. 1994) and the resulting alignments were modified manually. In the analysis, sites that include gaps were removed. For phylogenetic analysis, we used the neighbor-joining (NJ) method based on the number of nucleotide differences (the *p*-distances) as well as the maximum parsimony (MP) method implemented in PHYLIP 3.572 (Felsenstein 1993).

ML Method

The two-species ML method used here is essentially the same as that in Takahata and Satta (1997). One difference is that the present method implements multiple-hit corrections. The largest observed nucleotide difference among the six primates is about 10% at most, so that multiple-hit corrections were made by the Jukes and Cantor (1969) method (see also Nei and Kumar 2000, p. 23). The computer program is written in *Mathematica* (version 3.0; Wolfram Research Champaign IL) and is available on request.

Result

Substitution Rates of Intron Sequences

To examine whether the data set of 20 intron sequences (O'hUigin et al. 2002) is representative of the entire genome in terms of nucleotide divergences, we compared those of humans and chimpanzees with a collection of pairs of BES (Fujiyama et al. 2002). The collection consisted of 58,156 BES pairs, from which we chose 20 pairs at random. Repeating this subsampling 1000 times, we obtained the distributions of their mean and variance of nucleotide divergences (Fig. 1). The mean and variance in the 20 pairs of human and chimpanzee intron sequences are 0.0147 and 4.39 × 10^{-5}, respectively, and both are within the 90% confidence regions of the mean and variance distributions for the BES random subsamples. We therefore concluded that the 20 intron sequences could be regarded as representatives of the human and chimpanzee genome.

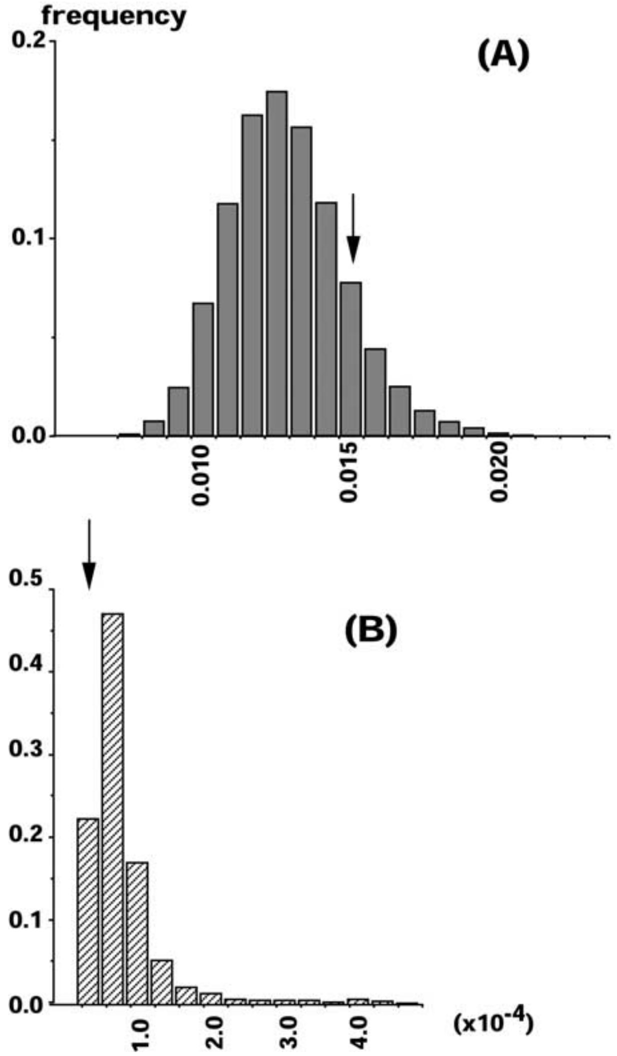

Fig. 1. Distribution of means (**A**) and variances (**B**) of nucleotide divergences in 20 resampled BES data. The distribution was obtained by 1000 replications. The arrow shows the class which contains the mean and variance in the 20 intron sequences, respectively. On the ordinate and abscissa are plotted the frequency and the range of mean or variance in the resampled data, respectively.

In comparisons between M or T and hominoids (H, C, G, or O), rate heterogeneity of nucleotide substitutions is apparent. O'hUigin et al (2002) showed that 10%–20% of substituted sites have experienced multiple hits, even when the average nucleotide divergence is as low as 10%. Multiple substitutions often result in phylogenetically incompatible sites within a single gene or region, and in the intron sequence data set, several phylogenetically incompatible sites are observed (Table 1). If the extent of this incompatibility differs greatly from locus to locus, the cause might be attributed to rate heterogeneity of nucleotide substitutions

Table 1. The number of phylogenetically informative sites and the number of nucleotides in introns of 20 loci (sequences from O'hUigin et al. 2002)

	Pattern of partition[a]						
Locus (bp)	(H,C)G	(H,G)C	(C,G)H	(H,C,G)O	(H,C,G,O)M	Incompatible sites (%)[b]	m_i/P_i[c]
ANP (120)	0	0	0	1	0	0 (0)	0.51/0.601
TNF (1052)	0	0	0	9	8	4 (0.38)	4.5/0.538
DAF (498)	0	0	0	3	7	3 (0.60)	2.1/0.836
HBBP (948)	0	0	0	9	15	5 (0.53)	4.0/0.376
B2M (721)	2	0	0	0	6	1 (0.14)	3.1/0.190
F9 (583)	3	0	0	2	10	2 (0.34)	2.5/0.550
PAH (1028)	2	0	0	4	14	3 (0.29)	4.4/0.365
LCAT (917)	1	0	0	7	9	5 (0.54)	3.9/0.351
BOP (940)	1	0	0	7	8	3 (0.32)	4.0/0.435
IL3 (668)	2	0	0	5	4	2 (0.30)	2.8/0.461
APOA1 (576)	0	0	3	9	4	1 (0.17)	2.4/0.298
UOX (1341)	0	0	3	5	22	3 (0.22)	5.7/0.181
C4B (404)	0	0	2	5	0	2 (0.50)	1.7/0.511
AFP (613)	0	3	0	2	6	2 (0.33)	2.6/0.518
EPO (932)	0	1	0	5	5	5 (0.54)	4.0/0.363
ODC1 (672)	0	1	1	8	8	2 (0.30)	2.9/0.457
GHR (1394)	1	0	2	3	13	5 (0.36)	5.9/0.459
HBE (917)	1	0	1	7	7	3 (0.33)	3.9/0.454
INS (669)	1	2	1	3	8	7 (1.10)	2.8/0.026
HBG (951)	1	2	1	7	7	12 (1.30)	4.0/0.001

[a] The number of sites supporting each of the five partitions is given. (H,C)G stands for partitions supporting the (H,C) cluster to the exclusion of G. Similarly, (H,C,G)O indicates the number of sites supporting the (H,C,G) cluster to the exclusion of O, and (H,C,G,O)M that of (H,C,G,O) to the exclusion of M.

[b] The number of incompatible sites. An incompatible site is defined as a site that requires more than one substitution to be compatible with the authentic phylogeny. Because the incompatibility observed among the members of the (H,C,G) trio is due not only to multiple substitutions, but also to recombination, this trio is excluded. The number in parentheses is the proportion (%) of the incompatible sites.

[c] At the *i*th locus, the probability of having an equal or larger (smaller) number of multiple hit sites than the observation is calculated according to the Poisson distribution with the mean of m_i. Taking the average (p_m) of the proportion of the multiple-hit sites across 20 loci, $m_i = p_m L_i$, where L_i is the number of sites compared at the *i*th locus.

among different loci. Therefore we counted the number of sites that experienced multiple substitutions by the maximum parsimony method, assuming the standard phylogenetic relationship among the six primate species, namely, ((((H,C,G)O)M)T). Incompatible sites among H, C, and G were ignored, because they have likely been generated by intragenic recombination (Satta et al. 2000; O'hUigin et al. 2002).

Interlocus rate heterogeneity was examined by the binomial distribution. Based on the average proportion of multiple hits over the 20 loci, the expected number (m_i) of multiple hits at the *i*th locus was calculated. We then obtained the probability (P_i) of having an equal or larger (smaller) number of multiple hits compared to the observation at the *i*th locus. Since the number of sites compared is large and m_i is small, we used the Poisson approximation to calculate P_i (Table 1). The result reveals that the insulin (*INS*) and γ-globin (*HBG*) introns show more frequent multiple substitutions than the expectation (Table 1; $p < 0.05$ and $p < 0.001$), suggesting that the nucleotide substitution rate at these loci is significantly higher than that at other loci.

Phylogenetic Relationships

NJ trees at 18 of 20 loci are topologically identical to the standard phylogenetic relationships of the six primates. The two exceptions are the β2 microglobulin (*B2M*) and complement 4B (*C4B*) loci. The *B2M* tree has no substitutions on a branch leading to a cluster of (H,C,G) and in the *C4B* tree the same thing happens on a branch leading to a cluster of (H,C,G,O). Examination of phylogeneticaly informative sites (Table 1) and MP analyses (data not shown) can confirm the absence of substitutions on these branches. However, since *B2M* and *C4B* do not show any significant shortages or excesses of the number of multiple hits (Table 1), it is unlikely that the unusual substitution patterns result from a slowdown or acceleration of the nucleotide substitution rate. Therefore we did not exclude these loci from the following ML analysis.

Intragenic Recombination Within Intron Sequences

To examine linkage between sites within a locus, we analyzed individual informative sites for their support of phylogenetic relationships among H, C, and G. The analysis reveals that four loci (*ANP*, *TNF*, *DAF*, and *HBBP1*) contain no informative sites with regard to the (H,C,G) relationships. Eleven loci contain informative sites that support one of the three possible relationships: *F9*, *B2M*, *PAH*, *LCAT*, *BOP*, and *IL3* support the (H,C)G; *APOA1*, *UOX*, and *C4B*, the (C,G)H; and *AFP* and *EPO*, the (H,G)C. Finally, the remaining five loci (*ODC1*, *GHR*, *INS*, *HBE*, and *HBG*) show that some sites support one relationship, while others favor a different one even at a single locus (Table 1). For example, the *HBG* locus contains four phylogenetically informative sites, one of which supports the (H,C)G, another the (C,G)H, and two the (H,G)C relationship. In these five loci, intergenic recombination is, therefore, likely to have occurred in the ancestral population of the three species. Since these five loci include *INS* and *HBG* at which loci rate heterogeneity is apparent, we may exclude them from the two-species ML analysis.

The Two-Species ML Method

The ancestral population size (N) and species divergence time (t_s) are obtained in terms of $x = 4Ng\mu$ and $y = 2t_s\mu$ in the two-species ML method. To check the reliability of these estimates for the 15 different pairs of the six primates, we divide these pairs into five classes with respect to shared ancestral populations. The classes are [(H,C)], [(H,G), (C,G)], [(H,O), (C,O), (G,O)], [(H,M), (C,M), (G,M), (O,M)], and [(H,T), (C,T), (G,T), (O,T), (M,T)]. We designate the ancestral populations of these classes HC, HCG, HCGO, HCGOM, and HCGOMT, respectively. By definition, members in each class share a common ancestral population immediately before their divergences. Thus, for example, the ancestral population of the (H,O) pair is also the ancestral population of the (G,O) and (C,O) pairs, and these three pairs are in turn all members of the same HCGO class. Because of the sharing of ancestral populations, the estimates of x and y must be the same for all the pairs in a given class, even though approximately.

Before excluding the five loci mentioned above, we applied the ML method to the entire data set of 20 intron sequences and estimated x and y for each of 15 pairs of species. The result reveals satisfactory consistency in estimates of y within each class and fairly large estimates of x, ranging from 0.42 to 1.5% (Table 2). With these as a reference, we applied the ML method to the trimmed data set, which excludes *ODC1*, *GHR*, *INS*, *HBE*, and *HBG* because of high nucleotide substitution rates or intragenic recombination (Table 1). The ML estimates of y are in good agreement with those for the entire data set. Because t_s remains constant among different loci, the y estimates are not much affected by trimming the data set. However, the x estimates become substantially small (Table 2).

Table 2. ML estimates (%) of $x = 4Ng\mu$[c] and $y = 2t_s\mu$[c] based on the entire or the truncated data set[b] of intron sequences and exon sequences

		Introns		Exons[e]	
Class[a]	Species pair	x[c]	y[c]	x	y
HC	H,C	0.33 (0.57)[d]	0.9 (0.9)	0.45	1.3
HCG	H,G	0.21 (0.54)	1.1 (0.9)	0.50	1.6
	C,G	0.22 (0.66)	1.1 (0.9)	n.a.	n.a.
HCGO	H,O	0.03 (0.59)	2.8 (2.5)	0.92	3.2
	C,O	0.08 (0.42)	2.9 (2.8)	n.a.	n.a.
	G,O	0.03 (0.80)	2.9 (2.5)	n.a.	n.a.
HCGOM	H,M	0.65 (0.66)	5.2 (5.3)	0.40	7.4
	C,M	0.78 (0.76)	5.1 (5.3)	n.a.	n.a.
	G,M	0.86 (0.92)	5.0 (5.1)	n.a.	n.a.
	O,M	0.81 (1.0)	5.1 (5.1)	n.a.	n.a.
HCGOMT	H,T	0.51 (1.0)	9.9 (9.6)	1.1	12
	C,T	0.80 (1.0)	9.3 (9.8)	n.a.	n.a.
	G,T	0.68 (1.1)	9.7 (9.5)	n.a.	n.a.
	O,T	0.41 (1.3)	10.0 (9.5)	n.a.	n.a.
	M,T	0.96 (1.5)	10.6 (10.3)	n.a.	n.a.

[a] Species pairs within each class shared an ancestral population.

[b] Excluded because of recombination or rate heterogeneity (see text).

[c] Where N is the effective ancestral population size of each pair of species, μ the nucleotide substitution rate per site per year, t_s the species divergence time, and g the generation time of an ancestral species.

[d] The numbers in parentheses are the ML estimates based on 20 loci.

[e] The estimates are taken from Takahata (2001), and n.a. means that the estimate is not available.

Discussion

Comparison Between Intron-Based and Exon-Based x *and* y *Estimates*

We compared the present estimates with the previous ML estimates based on exon sequences (Table 2). It is interesting that the x estimates based on exon sequences are close to those on the entire data set of intron sequences, suggesting that exon data still contain heterogeneous sequences regarding the nucleotide substitution rate or intragenic recombination. On the other hand, the y estimates based on exons are much larger than those on introns. Considering that the y estimates are not much affected by the exclusion of outrageous sequences (Table 2), the relatively large y estimates based on the exon sequences are caused by an overestimation of synonymous divergences, but not by rate heterogeneity of nucleotide substitutions.

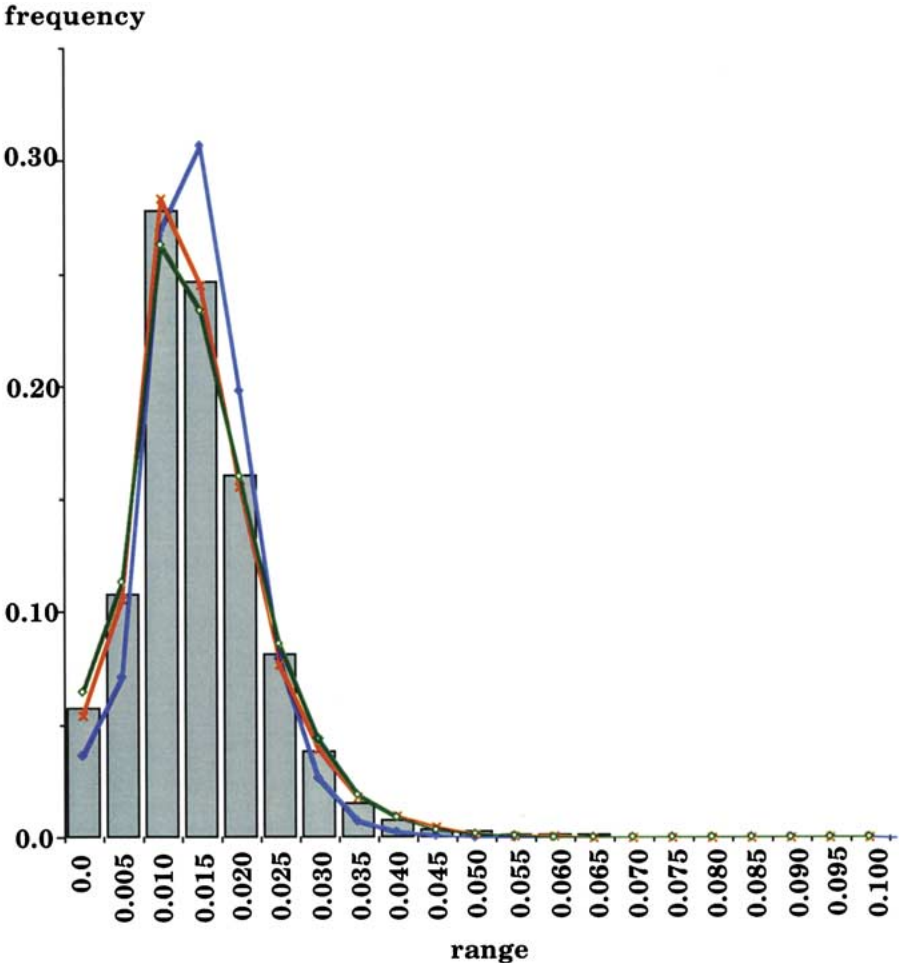

Fig. 2. Per-site difference distribution of BES data and of data obtained by simulations (generating random variables) under three models. The number of sites for each locus is the same as for BES. Bars indicate the observed results and lines represent the results of simulations, respectively. The blue line represents the Poisson distribution; the red line, the geometric + Poisson distribution; and the green line, the negative binomial distribution.

The x estimates are generally smaller in the trimmed data set than in the entire data set of 20 intron sequences as well as in the exon data set. In particular, the x estimate for the HCGO class is consistently much smaller than that of any other (Table 2) and is as small as that of extant humans (ca. 0.1%). Although further accumulation of intron sequences is necessary, this may suggest that the primate lineage has experienced a reduction of the population size when Asian apes diverged from African apes (see later).

Patterns of Nucleotide Substitutions in Human–Chimpanzee Comparisons

We examined whether or not the variation of the nucleotide divergences observed in human–chimpanzee comparisons can be explained by factors other than a relatively large ancestral population size. Specifically we focused on the effect of a limited number of sites compared and different substitution rates among different loci. To evaluate the effect, we performed a computer simulation that imitates the human and chimpanzee BES data.

We consider three nucleotide substitution models. The first focuses on the variation of the nucleotide divergence caused by a limited number of sites compared at individual BES loci. We use a constant nucleotide substitution rate and assume that the coalescence time in the ancestral population is negligibly small compared to the species divergence time. Using the observed mean nucleotide divergence per site (d_{HC}) over 58,158 BES loci, the expected number of nucleotide substitutions at the ith BES locus is estimated as $d_{\mathrm{HC}}L_i$, where L_i is the number of nucleotides compared. Setting $d_{\mathrm{HC}}L_i$ as a Poisson parameter, we generate a Poisson random variable (k_i) for the ith locus and calculate the number of nucleotide substitutions per site as k_i/L_i. Repeating this process 58,158 times, we obtain the distribution of k_i/L_i (blue line in Fig. 2).

The second model is based on the negative binomial distribution, and, as in the first, we ignore the presence of ancestral polymorphism. The variation of nucleotide divergences is then attributed mainly to the variation in the substitution rate among different loci (Yang 2002). Following equation (5.14) in Takahata and Satta (2002), we estimate the shape parameter α (α = 5.82) of the gamma distribution of

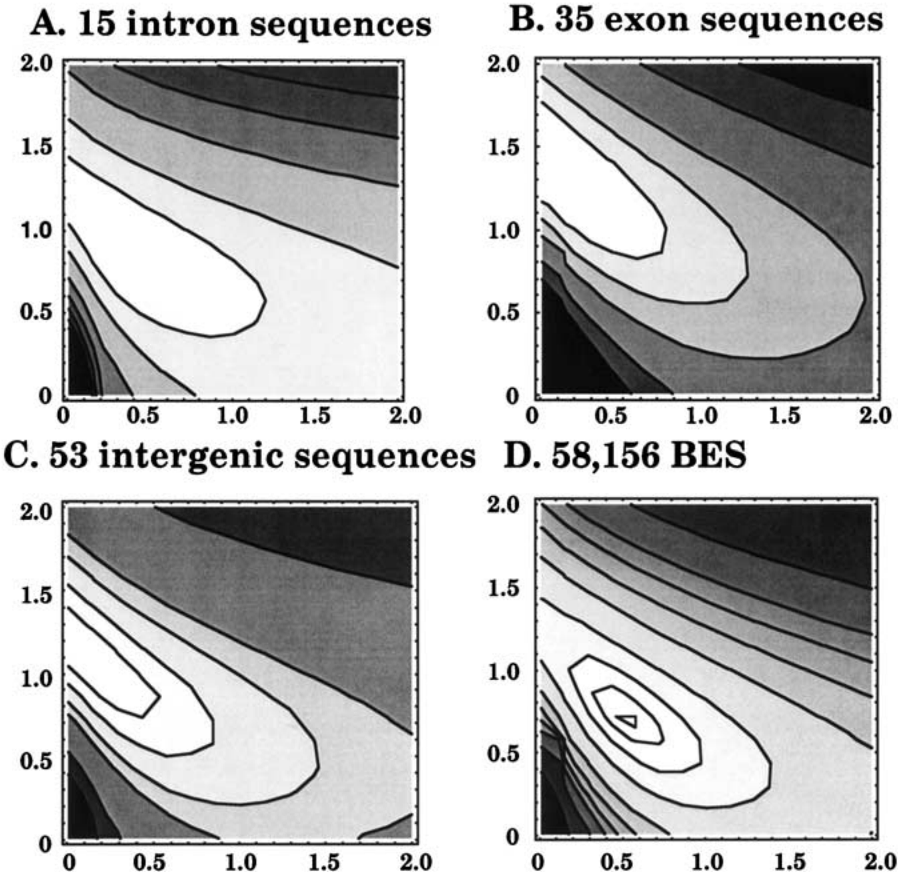

Fig. 3. Contour plots of the log likelihood function of four data sets which compare humans with chimpanzee nucleotide sequences: (**A**) 15 intron sequences, (**B**) 35 exon sequences, (**C**) 53 intergenic sequences, and (**D**) 58,156 BES. The abscissa and the ordinate give the range of the estimate of $x = 4Ng\mu$ and $y = 2t_s\mu$, respectively. The ML estimates for A–D are given in Table 3. The innermost area in A, B, and C represents the 90% confidence region of the x and y estimates. In D, the innermost area shows the 99.9% confidence region.

the nucleotide substitution rate and calculate the mean substitution rate (r) from $d_{HC} = 2rt_s$, assuming that $t_s = 6 \times 10^6$ years (Brunei et al. 2002). We then generate a gamma variable γ_i for the substitution rate at the ith locus and determine the number of nucleotide substitutions (k_i) by following the Poisson distribution with mean $2t_sL_i\gamma_i$. This procedure is equivalent to generating a random variable that follows the negative binomial distribution. Again repeating this process 58,156 times, we obtain the distribution of k_i/L_i (green line in Fig. 2).

The third model is based on the convolution of the geometric and Poisson distributions, as derived in Takahata et al. (1995), and takes explicit account of ancestral polymorphism. Some extent of the variation in the number of nucleotide substitutions can be attributed to the variation in coalescence times in the ancestral population. To simulate this model, we first estimate x_{HC} and y_{HC} from the 58,156 BES data. Since $y_{HC} = 2t_s\mu$, we generate a Poisson variable with mean $y_{HC}L_i$ for the number of nucleotide substitutions at the ith locus that can accumulate after the species divergence. We also generate a random variable that is geometrically distributed with mean $x_{HC}L_i$ for the number of nucleotide substitutions during the phase of ancestral polymorphism. Dividing the sum of these Poisson and geometric random numbers by L_i, we obtain a per-site random variable ($x_i + y_i$) for each of 58,156 loci and plot the distribution (red line in Fig. 2).

As expected, the mean of sequence divergences in each of the above three models is the same as the observation (0.0124). However, the variance varies depending on models. Whereas the variance in the third convolution model (7.54×10^{-5}) is in good agreement with the observation (7.55×10^{-5}), the variance in both the Poisson and the negative binomial models (4.22×10^{-5} and 6.85×10^{-5}) is somewhat small. In fact, the Kolmogorov–Smilnov test (Sokal and Rohlf 1969, pp. 704–721) reveals that the first and second models do not fit the observation ($p < 0.01$ for each case). We therefore conclude that the distribution of sequence divergences best fits the observation of the BES data set under the convolution model (Fig. 2; $p > 0.05$). We also find that at least for humans and chimpanzees, the variation in nucleotide divergences among loci does not appear to be much affected by heterogeneity in nucleotide substitution rates.

Human–Chimpanzee Ancestral Population Size

There are four data sets of nucleotide sequences, which can be used for the ML estimation of the ancestral human and chimpanzee population size (Chen and Li

Table 3. Estimates of the extent of polymorphism ($x = 4Ng\mu$) or effective size (N) in the ancestral population and of species divergence ($y = 2t_s\mu$) for humans and chimpanzees by the maximum likelihood (ML) method

	ML[a]			
	Chen and Li (2001)	Takahata (2001)	O'hUigin et al. (2002)	Fujiyama et al. (2002)
No. of loci	53	37	15	58,156
x (%)[b]	0.099 (0.100)[a]	0.45	0.33	0.51
y (%)[b]	1.04 (1.06)	1.32	0.90	0.73

[a]ML estimates in parentheses were obtained under the model taking rate variation into consideration (Yang 2002). ML estimates for O'hUigin et al. data are based on intron sequences only.
[b]Parameters for x and y are the same as in Table 2.

2001; Takahata 2001; O'hUigin et al. 2002; Fujiyama et al. 2002). They are 53 intergenic regions, 37 exonic regions, 15 introns, and 58,156 BES, respectively. Of these, the ML estimate from the BES data seems the most reliable because of an exceptionally large number of loci examined (Fig. 3). To evaluate the effect of the number of loci on our ML estimates, we resample 20, 50, or 100 loci from the BES data and examine the estimates of x and y based on 1000 such replications. The estimates obtained for the entire BES data are $x = 0.51\%$ and $y = 0.73\%$. However, as the number of loci becomes small, the range of both x and y estimates becomes broad. Even for 100 loci and under the condition of 95% confidence limits, the x estimate ranges from 0.25 to 0.76% and the y estimate ranges from 0.59 to 0.99% (data not shown). The 90% confidence region of x and y for BES is extremely small compared with that for other data sets (Fig. 3).

It may be noted that the x estimate for the intergenic sequences in Chen and Li (2001) is quite small (Table 3). To make a quantitative assessment, we calculate the mean and variance of nucleotide substitutions over the 53 loci and compare them with those in 1000 replications of 53 resampled BES data sets. The mean of Chen and Li's data set is 1.23%, which is in good agreement with the 1.24% for the resampled BES data. However, the variance of Chen and Li's data is only 3.01×10^{-5}, which is significantly smaller ($p < 0.01$) than that of the BES data (7.55×10^{-5}). Thus, although cause is unknown, Chen and Li's data show an unexpected uniformity in the extent of nucleotide substitutions between humans and chimpanzees.

Yang (2002) developed a method for estimating an ancestral population size using ML and Bayesian approaches. Taking into consideration different substitution rates among different loci, he applied these approaches to Chen and Li's data and obtained $x = 0.1\%$, which is almost the same as that for extant humans (Li and Sadler 1991). However, if the ancestral population size were the same as the extant human population size (10^4), most pairs of H and C orthologous genes should have coalesced within the ancestral population. Under the assumption of the ancestral population size of 10^4 individuals and the interval of $T = 1$ myr between the human–chimpanzee divergence and the (human–chimpanzee)–gorilla divergence, the proportion of discordance between the species and the gene tree becomes 0.1% from the trichotomy method (Nei 1987). In other words, 99.9% of the data should have supported the (H,C)G relationship, but in fact only 42% do (Chen and Li 2001). The small estimated value of x is not owing to the methodology since the simple two-species ML method also gives $x = 0.099\%$ (Table 3). Since the small estimate of x cannot be achieved by taking heterogeneity of nucleotide substitution rates, it must result from an unusual small variance in the number of substitutions.

There still remain differences among the ML estimates of x and y in other data sets (Table 3). Nonetheless, we can draw two conclusions. First, except for the estimate by Takahata (2001), which appears to be affected by overestimation of synonymous divergences, the y estimate for chimpanzees and humans ranges only from 0.73 to 1.04% (Table 3). Assuming the divergence time of 6 myr between the two species (Brunei et al. 2002), we estimate the nucleotide substitution rate as $0.6–0.8 \times 10^{-9}$ per site per year. This rate is lower than generally accepted (cf. Li 1997). Second, the x estimate ranges from 0.33 to 0.51%. These values are four to five times larger than the estimate of the extant human population (Li and Sadler 1991). If we further take account of a prolonged generation time of extant humans, the effective size of the ancestral human–chimpanzee population must have been approximately 10 times larger than 10^4 for extant humans (Takahata and Satta 1997; Takahata 2001).

Demographic History of Primate Populations During the Last 50 myr

Discrepancies between molecular and paleontological estimates of primate divergence time have been pointed out recently (Martin 1993; Tavaré et al. 2002), and a new statistical approach pushes the last common ancestor of primates back as old as 81.5 myr ago. Martin (1993) suggested that the divergence time of major nodes in the primate phylogeny was pushed back at least 10 myr, and our results support this view. If we assume that the divergence time between humans and chimpanzees is 6 myr (Burnet et al. 2002), our ML estimates of y (Table 2) suggest that the divergence times of the major nodes in the primate phylogeny become 7.2 myr for (H,C)G, 18 myr for (H,C,G)O, 34 myr for (H,C,G,O)M, and 65 myr for (H,C,G,O,M)T. These divergence times are older than that indicated by fossil records.

Recently, there are several molecular approaches to estimate the divergence time of primate species (Kumar and Hedges 1998; Glazko and Nei 2003; Hasegawa et al. 2003). When we compare our results with these estimates, our estimate of the divergence time of gorillas from humans (7.2 myr) shows good agreement with others (ranging 7 to 12 myr). Similarly, our estimate of the divergence time of humans from orangutan (18 myr) appears to be in the range of others (ranging 8 to 18 myr). In addition, this relatively old divergence time of orangutans is consistent with the time when the African continent became combined with Eurasia some 18 myr ago (Waddell and Penny 1996). However, regarding more ancient divergences, there are large discrepancies among various estimates. For instance, Kumar and Hedges (1998) estimated the divergence time of OWMs from humans as 21–24 myr and Glazko and Nei (2003) obtained a similar estimate (21–25 myr). On the other hand, Hasegawa et al. (2003) estimated the divergence to be as old as 31–38 myr. Our estimate is consistent with the latter. Furthermore, Kumar and Hedges (1998) and Glazko and Nei (2003) estimated the date of NWM divergence as 39–56 and 32–36 myr, respectively. On the other hand, our estimate was much older (65 myr). Although this discrepancy may come from different data and methods used, it is evident that more studies for the primate phylogeny are necessary, especially to reach a consensus about the divergence time.

To convert the amount of ancestral polymorphism (measured by $x = 4Ng\mu$) into the effective size (N) of the ancestral population, information on the generation time in that population is required. Although there are uncertainties about the generation time of nonhuman primates, it is shorter than the generation time of extant humans (Gavan 1953). Under this assumption, the estimated values of x suggest that the ancestral population size has been of the order of 10^5 throughout most of primate evolution, although there might be an occasional reduction as discussed earlier. It also appears that such a large size of the ancestral population of humans, chimpanzees, and gorillas is consistent with the high extent of DNA polymorphism in extant nonhuman primates (Kaessmann et al. 1999, 2001; Satta 2001).

Acknowledgments. We thank N. Takahata for his critical reading of an early version of this paper. We also thank two anonymous reviewers for their helpful comments. This research was supported in part by Japan Society for Promotion of Science Grant 12304046 and in part by a MEXT summer program grant to M.H.

References

Bielawski JP, Dunn KA, Yang Z (2000) Rates of nucleotide substitution and mammalian nuclear gene evolution: Approximate and maximum-likelihood methods lead to different conclusions. Genetics 156:1299–1308

Brunet M, Guy F, Pilbeam D, et al. (2002) A new hominid from the Upper Miocene of Chad, Central Africa. Nature 418:145–151

Chen F-C, Li W-H (2001) Genomic divergence between human and other hominoids and the effective population size of the common ancestor of human and chimpanzee. Am J Hum Genet 68:444–456

Edwards SV, Beerli P (2000) Perspective: Gene divergence, population divergence, and the variance in coalescence time in phylogeographic studies. Evolution 54:1839–1854

Felsenstein J (1993) PHYLIP (Phylogeny Inference Package) version 3.572. Distributed by the author, Department of Genetics, University of Washington, Seattle

Fujiyama A, Watanabe H, Toyoda A, et al. (2002) Construction and analysis of a human-chimpanzee comparative clone map. Science 295:131–134

Gavan JA (1953) Growth and development of the chimpanzee; A longitudinal and comparative study. Hum Biol 25:93–143

Glazko GV, Nei M (2003) Estimation of divergence times for major lineages of primate species. Mol Biol Evol 20:424–434

Goodman M, Porter CA, Czelusniak J, Page SL, Schneider H, Shoshani J, Gunnell G, Groves CP (1998) Toward a phylogenetic classification of primates based on DNA evidence complemented by fossil evidence. Mol Phylogenet Evol 9:585–598

Hartl DL, Clark AG (1997) Principles of population genetics, 3rd ed. Sinauer Associates, Sunderland, MA

Hasegawa M, Thorne JL, Kishino H (2003) Time scale of eutherian evolution estimated without assuming a constant rate of molecular evolution. Genes Genet Syst 78:267–283

Jukes TH, Cantor CR (1969) Evolution of protein molecules. In: Munro HN (ed) Mammalian protein metabolism III. Academic Press, New York, pp 21–132

Kaessmann H, Wiebe V, Pääbo S (1999) Extensive nuclear DNA sequence diversity among chimpanzees. Science 286: 1159–1162

Kaessmann H, Wiebe V, Weiss G, Pääbo S (2001) Great ape DNA sequences reveal a reduced diversity and an expansion in humans. Nature Genet 27:155–156

Kimura M (1983) The neutral theory of molecular evolution. Cambridge University Press, Cambridge

Kingman JFC (1982) On the genealogy of large populations. J Appl Prob A19:27–43

Kumar S, Hedges B (1998) A molecular timescale for vertebrate evolution. Nature 392:917–920

Li W-H (1997) Molecular evolution. Sinauer Associates, Sunderland, MA

Li W-H, Sadler LA (1991) Low nucleotide diversity in man. Genetics 129:513–523

Martin RD (1993) Primate origins: Plugging the gaps. Nature 363:223–234

Nei M (1987) Molecular evolutionary genetics. Columbia University Press, New York

Nei M, Kumar S (2000) Molecular evolution and hylogenetics. Oxford University Press, New York

O'hUigin C, Satta Y, Takahata N, Klein J (2002) Contribution of homoplasy and of ancestral polymorphism to the evolution of genes in anthropoid primates. Mol Biol Evol 19:1501–1513

Pilbeam D (1984) The descent of hominoids and hominids. Sci Am 250:84–96

Rannala B, Yang Z (2003) Bayes estimation of species divergence times and ancestral population sizes using DNA sequences from multiple loci. Genetics 164:1645–1656

Satta Y (2001) Comparison of DNA and protein polymorphisms between humans and chimpanzees. Genes Genet Syst 76:159–168

Satta Y, Klein J, Takahata N (2000) DNA archives and our nearest relative: the trichotomy problem revisited. Mol Phylogenet Evol 14:259–275

Sokal RR, Rohlf FJ (1969) Biometry. W.H. Freeman, New York

Tajima F (1983) Evolutionary relationship of DNA sequences in finite populations. Genetics 105:437–460

Takahata N (2001) Molecular phylogeny and demographic history of humans. In: Tobias PV, Taath MA, Moggi-Cecchi J, Doyle GA (eds) Humanity from African naissance to coming millennia. Firenze University Press, Johannesburg, pp 299–305

Takahata N, Satta Y (1997) Evolution of the primate lineage leading to modern humans: Phylogenetic and demographic inferences from DNA sequences. Proc Natl Acad Sci USA 94:4811–4815

Takahata N, Satta Y (2002) Pre-speciation coalescence and the effective size of ancestral populations. In: Slatkin M, Veuille M (eds) Modern developments in theoretical population genetics. Oxford University Press, New York, pp 52–71

Takahata N, Satta Y, Klein J (1995) Divergence time and population size in the lineage leading to modern humans. Theor Popul Biol 48:198–221

Tavaré S, Marshall CR, Will O, Soligo C, Martin RD (2002) Using the fossil record to estimate the age of the last common ancestor of extant primates. Nature 416:726–729

Thompson JD, Higgins DG, Gibson TJ (1994) CLUSTAL W: Improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res 22:4673–4680

Waddell P, Penny D (1996) Evolutionary trees of apes and humans from DNA sequences. In: Lock AJ, Peters CR (eds) Handbook of human symbolic evolution. Oxford University Press, Oxford, pp 53–73

Wall JD (2003) Estimating ancestral population sizes and divergence times. Genetics 163:395–404

Wu C-I (1991) Inferences of species phylogeny in relation to segregation of ancient polymorphisms. Genetics 127:429–435

Yang Z (1997) On the estimation of ancestral population sizes of modern humans. Genet Res Cambr 69:111–116

Yang Z (2002) Likelihood and Bayesian estimation of ancestral population sizes in hominoids using data from multiple loci. Genetics 162:1811–1823