Estimation of Hominoid Ancestral Population Sizes under Bayesian Coalescent Models Incorporating Mutation Rate Variation and Sequencing Errors

Ralph Burgess† and Ziheng Yang*‡*

*Galton Laboratory, Department of Biology, University College London, London, United Kingdom; †Watson School of Biological Sciences, Cold Spring Harbor Laboratory, Cold Spring Harbor, NY; and ‡Laboratory of Biometrics, Graduate School of Agriculture and Life Sciences, University of Tokyo, Tokyo, Japan

Estimation of population parameters for the common ancestors of humans and the great apes is important in understanding our evolutionary history. In particular, inference of population size for the human–chimpanzee common ancestor may shed light on the process by which the 2 species separated and on whether the human population experienced a severe size reduction in its early evolutionary history. In this study, the Bayesian method of ancestral inference of Rannala and Yang (2003. Bayes estimation of species divergence times and ancestral population sizes using DNA sequences from multiple loci. Genetics. 164:1645–1656) was extended to accommodate variable mutation rates among loci and random species-specific sequencing errors. The model was applied to analyze a genome-wide data set of ~15,000 neutral loci (7.4 Mb) aligned for human, chimpanzee, gorilla, orangutan, and macaque. We obtained robust and precise estimates for effective population sizes along the hominoid lineage extending back ~30 Myr to the cercopithecoid divergence. The results showed that ancestral populations were 5–10 times larger than modern humans along the entire hominoid lineage. The estimates were robust to the priors used and to model assumptions about recombination. The unusually low X chromosome divergence between human and chimpanzee could not be explained by variation in the male mutation bias or by current models of hybridization and introgression. Instead, our parameter estimates were consistent with a simple instantaneous process for human–chimpanzee speciation but showed a major reduction in X chromosome effective population size peculiar to the human–chimpanzee common ancestor, possibly due to selective sweeps on the X prior to separation of the 2 species.

Introduction

The effective size N of a population is directly related to the genetic diversity that is maintained in the population (Kimura and Crow 1964). Analyses of polymorphism data have consistently estimated $N \approx 10{,}000$ for the modern human lineage (e.g., Nei and Graur 1984; Takahata 1993; Yu et al. 2001). Furthermore, a study of genome-wide linkage disequilibrium, using an independent method, suggested that our recent population size was 2–3 times smaller (Tenesa et al. 2007). Although it is well recognized that effective population size may be much smaller than census population size, these estimates are still surprisingly low, and there is considerable interest in determining possible population bottlenecks in human evolutionary history. As for other modern hominoids, studies of polymorphism in nuclear noncoding regions estimated chimpanzee population size to be $N \approx 21{,}000$ and gorilla $N \approx 25{,}000$ (Yu et al. 2001, 2004; see also Kaessmann et al. 2001). Limited data for the orangutan suggested higher diversity, but an excess of intermediate-frequency alleles implicated population subdivision rather than greater size (Fischer et al. 2006).

Under the Fisher–Wright model, the expected time for a large sample of alleles at a neutral locus to find their most recent common ancestor is ~4N generations, with standard deviation (SD) at ~2.15N (e.g., Tajima 1983). Therefore, with a generation time of ~20 years and an effective population size $N \approx 10{,}000$, human diversity at neutral loci contains no demographic information much beyond 1.5 Myr. However, at the major histocompatibility complex (MHC), balancing selection has extended the mean coalescence time for human *DRB1* alleles to ~29 Myr (Satta et al. 1991). Takahata (1993) has argued that the diversity and deep coalescence of MHC alleles imply $N \approx 100{,}000$ over this longer timescale.

A more ancient demographic history of the human lineage may be inferred through comparison with the genomes of other primates. The evolutionary distance between orthologous sequences from 2 species, such as human (H) and chimpanzee (C), is attributable to 2 time components: the speciation time τ_{HC}, which is common to all loci, and the coalescent time in the HC ancestral population, which is variable among loci. Calendar times are scaled into evolutionary distances, with $\tau_{HC} = T_{HC}\mu$, where T_{HC} is the time in years to human–chimpanzee speciation and μ is the mutation rate, measured by the number of mutations per site per year. Coalescence rate and variance are determined solely by the scaled ancestral population size $\theta_{HC} = 4N_{HC}g\mu$, where g is the generation time in years. With a single locus, one cannot distinguish recent speciation with a large ancestral population (small τ_{HC} and large θ_{HC}) from ancient speciation with a small ancestral population (large τ_{HC} and small θ_{HC}). However, Takahata (1986) pointed out that data from multiple loci contain information about θ_{HC} in the *variation* of observed divergences among loci, enabling us to estimate τ_{HC} and θ_{HC} jointly. Such inference is nevertheless sensitive to mutation rate variation among loci (Yang 1997b).

With data for 3 or more closely related species, additional information is provided by incomplete lineage sorting. Consider the case of human (H), chimpanzee (C), and gorilla (G), where the species tree is ((HC)G) (see fig. 1). Coalescence of H and C alleles may occur in the internode ancestral population HC, in which case the gene tree matches the species tree. Otherwise, if all 3 alleles enter the HCG ancestral population, they may coalesce in any order with equal probability, with 2 of the 3 possible gene trees in conflict with the species tree. The probability of species tree–gene tree mismatch (P_{SG}) is thus two-thirds the

Key words: hominoid, ancestral population size, Bayesian inference, MCMC, coalescent model, sequencing errors.

E-mail: z.yang@ucl.ac.uk.

Mol. Biol. Evol. 25(9):1979–1994. 2008
doi:10.1093/molbev/msn148
Advance Access publication July 4, 2008

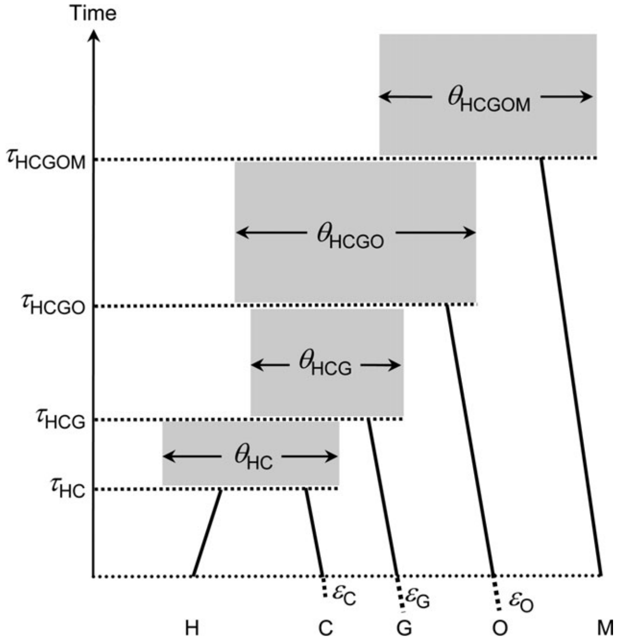

FIG. 1.—The species tree for human (H), chimpanzee (C), gorilla (G), orangutan (O), and macaque (M) showing the parameters in the model. For each ancestral population (referred to as HC, HCG, HCGO, HCGOM), 2 parameters are defined: $\theta = 4N\mu g$ and $\tau = T\mu$, where N is the effective population size, μ is the mutation rate per site per year, g is the generation time, and T is the species divergence time. Sequencing errors in C, G, and O are modeled by excess branch lengths ε_C, ε_G, and ε_O.

probability that no coalescence occurs in the ancestral population HC

$$P_{SG} = \frac{2}{3} e^{-2(\tau_{HCG} - \tau_{HC})/\theta_{HC}} \tag{1}$$

(Hudson 1983). This probability is greater for larger θ_{HC} and shorter internode time ($\tau_{HCG} - \tau_{HC}$). In real data sets, the true gene trees are unknown, and a common strategy is the so-called "tree-mismatch" or "trichotomy" method (e.g., Chen and Li 2001), by which the theoretical probability (P_{SG}) is equated to the proportion (P_{SE}) of mismatches between the species tree and the *estimated* gene tree. However, Yang (2002) pointed out that errors in gene tree reconstruction inflate the mismatch probability so that P_{SE} is always greater than P_{SG}, and this method may seriously overestimate ancestral population size θ_{HC} (see also below).

An alternative to the tree-mismatch method is the full likelihood approach, including both maximum likelihood (Takahata et al. 1995; Yang 2002) and Bayesian methods (Rannala and Yang 2003). The likelihood function accommodates uncertainty in the gene trees, weighting each possible tree by its probability of occurrence. Another advantage of the likelihood-based approach is that it takes account of the branch lengths in the trees, which provide information about gene coalescent times. The maximum likelihood calculation involves multidimensional integrals and is practical for small data sets only. The Bayesian Markov chain Monte Carlo (MCMC) algorithm, as implemented in the MCMCCOAL program (Yang 2002; Rannala and Yang 2003), was feasible for analyzing large data sets from multiple species, but few comparative primate data sets were available in 2003.

In this study, we used MCMCCOAL for Bayesian coalescent analysis of a genome-wide data set of ~15,000 neutral loci (7.4 Mb) from 5 primate species (Patterson et al. 2006). We updated the alignments to incorporate the high-quality genome assembly sequence now available for chimpanzee and macaque and trimmed the error-prone ends of shotgun sequencing reads, leading to improved data quality. The basic model in MCMCCOAL (Rannala and Yang 2003) was extended to accommodate variable mutation rates among loci and to allow for species-specific random sequencing errors. We discuss the implications of our parameter estimates for the speciation process between human and chimpanzee.

Materials and Methods

Sequence Data

We retrieved the alignments of human, chimpanzee, gorilla, orangutan, and macaque (HCGOM) of Patterson et al. (2006). Those data were gorilla whole-genome shotgun reads aligned to human assembly sequence, to multiple preassembly reads of chimpanzee and macaque, and to low-coverage orangutan reads. We combined overlapping reads into a single consensus sequence for each species and extracted 50,321 segments with at least 300 sites aligned for all species present in the alignment (table 1, row a). With lower read coverage for orangutan in the public databases, 38% of segments included only the 4 species HCGM. In most cases, segment ends corresponded to the end of a gorilla shotgun read, or less often an orangutan read, with mean length ~700 sites.

We improved the quality of the sequence data by incorporating the most recent genome assembly data for chimpanzee (*panTro2*, WashU build 2.1, October 2005) and macaque (*rheMac2*, Baylor build 1.0, January 2006). Human loci were identified in the University of California, Santa Cruz (UCSC) Genome Bioinformatics pairwise BlastZ alignments of the most recent human assembly *hg18* (NCBI build 36.1, March 2006) against chimpanzee and macaque. Four percent of the alignments were discarded because the loci mapped to breaks in the UCSC alignments for HC or HM. The sequences were realigned using MUSCLE (Edgar 2004). Default settings were used, except for a reduced gap-opening penalty of 250, found empirically to reduce the number of poor alignments. In a small number of cases (144, 0.3%), substituting UCSC-aligned chimpanzee and macaque sequences and realigning resulted in more than doubling of at least one pairwise distance. These loci were discarded as either the original alignments or the UCSC alignments may have been spurious.

Alignment ends corresponded to the error-prone ends of G or O shotgun sequencing reads (see Results), so they were truncated by 100 sites. (Truncating more sites showed negligible further improvement in quality.) This approach was preferred to the use of PHRED scores for individual base calls because our model required continuous alignments (without internal gaps). Furthermore, the use of PHRED thresholds may introduce bias (Johnson and Slatkin 2008), and they do not account for all sources of sequencing error. We preferred a conservative curation

Table 1
Data Set Statistics

Data Set	Number of Loci	Mean Size (bp)	Pairwise JC69 Distances										Kimura's κ			
			HC	HG	CG	HO	CO	GO	HM	CM	GM	OM	HC	HG	HO	HM
Data refinement																
(a) Uncurated raw alignments	50,321	702	0.0136	0.0195	0.0210	0.0353	0.0369	0.0394	0.0635	0.0652	0.0679	0.0672	3.6	3.4	3.8	4.0
(b) Incorporate C, M assembly	48,467	694	0.0122	0.0187	0.0189	0.0345	0.0347	0.0380	0.0613	0.0614	0.0648	0.0644	4.3	3.7	4.0	4.4
(c) 2 ×100-bp ends (removed)		200	0.0123	0.0248	0.0251	0.0373	0.0375	0.0462	0.0613	0.0616	0.0706	0.0671	4.2	3.0	3.7	4.4
(d) *Complete*	48,467	494	0.0121	0.0163	0.0165	0.0334	0.0336	0.0347	0.0612	0.0614	0.0625	0.0632	4.3	4.3	4.2	4.4
Neutral	**14,663**	**508**	**0.0127**	**0.0168**	**0.0170**	**0.0346**	**0.0348**	**0.0359**	**0.0625**	**0.0627**	**0.0638**	**0.0648**	**4.2**	**4.3**	**4.2**	**4.3**
Inert TE	**19,192**	**321**	**0.0128**	**0.0171**	**0.0174**	**0.0357**	**0.0360**	**0.0372**	**0.0664**	**0.0667**	**0.0677**	**0.0688**	**4.4**	**4.4**	**4.3**	**4.5**
LINE	9,877	313	0.0120	0.0159	0.0162	0.0331	0.0334	0.0344	0.0611	0.0614	0.0623	0.0636	4.1	4.2	4.1	4.1
SINE	9,041	215	0.0142	0.0192	0.0194	0.0406	0.0408	0.0421	0.0762	0.0766	0.0777	0.0785	5.1	4.9	4.9	5.4
LTR retrotransposon	4,213	292	0.0131	0.0174	0.0177	0.0354	0.0358	0.0371	0.0657	0.0660	0.0673	0.0680	4.2	4.4	4.2	4.4
DNA transposon	2,810	237	0.0124	0.0165	0.0166	0.0338	0.0340	0.0352	0.0633	0.0637	0.0646	0.0655	4.4	4.3	4.2	4.3
Neutral X	**783**	**477**	**0.0088**	**0.0137**	**0.0138**	**0.0278**	**0.0277**	**0.0286**	**0.0529**	**0.0529**	**0.0538**	**0.0548**	**4.2**	**4.1**	**4.0**	**4.2**
$d_{(X)}/d_{(A)}$			0.70	0.82	0.81	0.80	0.80	0.80	0.85	0.84	0.84	0.84				

Results for the principal data sets are highlighted in bold.

to avoid the risk of bias or overfiltering, which may remove genuinely variable sites. Our data set statistics (see Results) showed that we obtained high-quality continuous alignments of ~500 bp. Residual sequencing errors were dealt with under the extended model.

Complete was the refined inclusive autosomal data set, from which 2 data sets of neutrally evolving loci were compiled. The first was called *Neutral*. Starting with *Complete*, we removed loci within 1 kb of UCSC known genes (Hsu et al. 2006). We also removed some loci identified by RepeatMasker (Smit AFA, Hubley R, and Green P, unpublished data), including noncoding RNA genes, simple sequence repeats, and low-complexity regions. However, we masked only a small subset of transposable elements (TEs) in families that were potentially active in the past 40 Myr (see below). We also masked segmental duplications identified by the UCSC browser table for *hg18*. Remaining loci were uncharacterized but assumed to be neutrally evolving. From a separate curated data set *Complete X*, the same procedures were applied to X-linked loci to compile the *Neutral X* data set.

The second autosomal data set was called *Inert TE*. Here, in contrast to the conventional masking approach (above) that assumes uncharacterized loci are neutral, we positively identified well-characterized TE loci with phylogenetic evidence for neutrality. A large proportion of primate genomes is recognizable transposons, most of which have been assigned to well-characterized families for which detailed dated phylogenies are available (see below). We selected elements, identified by RepeatMasker, known to have been inert evolutionary fossils for at least 40 Myr. In practice, there was considerable overlap between the *Neutral* and *Inert TE* data sets. The major differences were that *Inert TE* included intronic transposons that were absent from *Neutral*, and that *Neutral* included uncharacterized intergenic unique sequence. *Inert TE* was partitioned into the 4 major primate TE classes as follows:

Long interspersed nuclear elements (LINEs): The young LINE families L1H, L1P1-3, and L1PA1-7 were recently active and were removed; older L1Ps and all L1Ms have been inert since Old World monkey (OWM) divergence and were included (Khan et al. 2006). The Lyon repeat hypothesis (Lyon 1998) proposed a functional role for LINEs on the X chromosome, but exclusion of X-enriched L1 elements (Bailey et al. 2000) from our data made little difference to the results, and they were included in the results shown.

Short interspersed nuclear elements (SINEs): Progressive enrichment of Alu SINEs in GC-rich regions has led to speculation about a functional role (Lander et al. 2001), but the absence of genetic variation even in young elements appears to rule out allelic selection; purifying selection against Alu–Alu ectopic recombination is the probable cause of enrichment (Batzer and Deininger 2002). Accordingly, only the young AluY family was excluded; AluJ, Sx1, and Sg1 were included, as were mammalian interspersed repeat (MIR) (Lander et al. 2001) and the Alu monomers (Quentin 1992).

Long terminal repeat (LTR) retrotransposons: The endogenous retroviruses (ERVs) probably derive from retroviruses, which lost infectivity but retained the ability to transpose within the host genome. However, the evolutionary dynamics of moderate and low-copy ERV families are often uncertain and may include reinfection (Bannert and Kurth 2006). Accordingly, we adopted a conservative approach, including only the well-characterized ancient nonautonomous mammalian LTR retrotransposons (MaLR) and LTR class III (or ERV-L), which are inert in higher primates (Smit 1993; Lander et al. 2001).

DNA transposons: There is no evidence of DNA transposon activity in the past 50 Myr, so all were included (Lander et al. 2001).

An assumption of the model is free recombination between loci. There is at present no definitive biological model for patterns of recombination over evolutionary time. Recombination appears to be concentrated at poorly characterized hot spots, which are not conserved even between humans and chimpanzees (Ptak et al. 2005). We calculated an order-of-magnitude estimate of the recombination rate by considering the total length of the human genome linkage map, ~3000 cM (Kong et al. 2004), across the 3×10^9 bp genome, implying a mean rate of $c \approx 10^{-8}$ per base pair per generation. With a generation time of 15 years, this amounts to 1 crossover per 1.5 kb per Myr. Thus, a minimum separation of 10 kb between loci seemed adequate to approximate the assumption of free recombination over the timescale of hominoid and OWM evolution. Loci were culled accordingly as the final step in preparation of the *Neutral* and *Inert TE* data sets.

Bayesian Inference

We use the Bayesian method of Rannala and Yang (2003; see also Yang 2002), implemented in the MCMCCOAL program. The method can be used to analyze sequence data from multiple loci from several closely related species, accounting for the species phylogeny and random coalescent events in extant and ancestral species. The JC69 mutation model (Jukes and Cantor 1969) is used for its computational efficiency and for the well-known robustness of analysis of highly similar sequences to the assumed mutation model. Here the role of the mutation model is to correct for multiple hits to estimate the gene tree topology and branch lengths. In previous studies, parsimony and neighbor-joining produced similar trees, and within the apes, even the infinite sites model produced very similar estimates to the finite-sites model (Satta et al. 2004). More complex models such as HKY+Γ (Hasegawa et al. 1985; Yang 1994) are thus deemed unnecessary in such analysis. The basic model assumes neutral evolution at a constant mutation rate, free recombination between loci, and no recombination within a locus. Several extensions to the basic model are introduced in this study.

Model of Sequencing Errors and Violation of the Molecular Clock

The human sequences are assumed to be error free. Sequencing errors in C, G, and O are modeled by adding ε_C, ε_G, and ε_O to the lengths of branches leading to C, G, and O in the gene trees (fig. 1). As the branch length is measured as the number of changes per site, the εs here represent the error rate per base pair. Each of ε_C, ε_G, and ε_O is assigned the gamma prior $G(1, 1000)$, with mean 0.0010 and 95% credibility interval (CI) 0.0003–0.0037. Note that the gamma distribution $G(\alpha, \beta)$ has mean α/β and variance α/β^2. Posterior distributions of ε_C, ε_G, and ε_O are generated by the MCMC algorithm. Because accommodating errors in M (by the use of parameter ε_M) would create an identifiability problem, ε_M is not estimated. In some analyses, ε_M is assigned a fixed value to model a higher mutation rate specific to the macaque lineage, as proposed in the hominoid slowdown hypothesis.

Our model assumes that errors affect sites and loci at random, at the same rate for the whole species. If the error rate varies over genomic regions (e.g., due to different sequencing coverage), it may be more realistic to let ε vary among loci according to a prior. This is not pursued here.

Variable Mutation Rates among Loci

We implement 2 methods to accommodate variable mutation rates among loci. In the first, *fixed-rates* model, we estimate the relative mutation rate for each locus by the average JC69 distances between macaque and the 4 apes:

$$d_{\text{HCGO-M}} = (d_{\text{HM}} + d_{\text{CM}} + d_{\text{GM}} + d_{\text{OM}})/4. \qquad (2)$$

This average distance is scaled such that the mean across all loci is 1, and the resulting relative rates are used to analyze data from 4 species only (HCGO) (Yang 2002).

The second model, referred to as the *random-rates* model, assumes random rate variation among loci. Let the rate for locus i be r_i, with $i = 1, 2, \ldots, L$. To avoid overparameterization, the average rate is fixed at one: $\bar{r} = \sum_{i=1}^{L} r_i / L = 1$. Parameters θs and τs are then defined using $\bar{r}$. We assign a Dirichlet prior on the transformed variables $y_i = r_i/L$, $i = 1, 2, \ldots, L$.

$$f(y_1, y_2, \ldots, y_L|\alpha) = \frac{\Gamma(L\alpha)}{[\Gamma(\alpha)]^L} \prod_{i=1}^{L} y_i^{\alpha-1}, \quad y_i > 0, \quad \sum_{i=1}^{L} y_i = 1. \qquad (3)$$

The marginal mean and variance are $E(y_i) = 1/L$ and $\text{var}(y_i) = (L-1)/[L^2(L\alpha+1)]$. Thus, the prior density on the rates, $r_i = Ly_i$, $i = 1, 2, \ldots, L$, is

$$\begin{aligned} f(r_1, r_2, \ldots, r_L|\alpha) &= \tfrac{\Gamma(L\alpha)}{\Gamma(\alpha)^L} \prod_{i=1}^{L} (r_i/L)^{\alpha-1} \times \tfrac{1}{L^L} \\ &= \frac{\Gamma(L\alpha)}{[L^\alpha \Gamma(\alpha)]^L} \prod_{i=1}^{L} r_i^{\alpha-1}, \quad r_i > 0, \\ &\quad \sum_{i=1}^{L} r_i = L. \end{aligned} \qquad (4)$$

We have $E(r_i) = 1$, $\text{var}(r_i) = (L-1)/(L\alpha+1) \approx 1/\alpha$ for large L, and $\text{corr}(r_i, r_j) = 1/(L-1)$. Parameter α is thus inversely related to the extent of rate variation among loci, and the impact of the prior can be assessed by changing α.

Note that both equations (3) and (4) are ($L - 1$)-dimensional densities. The first ($L - 1$) rates are working variables in the MCMC, with the last rate $r_L = L - \sum_{i=1}^{L-1} r_i$. A sliding window is used to update r_i, $i = 1, 2, \ldots, L - 1$. The new rate is generated as $r_i^* \sim U\left(r_i - \varepsilon/2, r_i + \varepsilon/2\right)$, reflected into the feasible range (0, $r_i + r_L$) if necessary, with $r_L^* = r_L - \left(r_i^* - r_i\right)$. The proposal ratio is one. The prior ratio of the move is, according to equation (4),

$$\frac{f\left(r_1, r_2, \ldots, r_i^*, \ldots, r_L^* | \alpha\right)}{f(r_1, r_2, \ldots, r_i, \ldots, r_L | \alpha)} = \left(\frac{r_i^* r_L^*}{r_i r_L}\right)^{\alpha - 1}. \qquad (5)$$

Implementation of Bayesian MCMC Algorithm

For each analysis, we ran the MCMCCOAL algorithm at least 3 times to confirm convergence. The MCMC was found to have good convergence and mixing properties. For analyses under different models, we used a burn-in of 5,000 iterations and then took 25,000 samples at every second iteration. The SD across the 3 runs was <0.5% of the posterior mean of the same parameter. For the major results, we ran longer chains, with 10,000 iterations for burn-in then 100,000 samples. There was no apparent difference between the long and short runs.

Results

Data Compilation and Quality

Table 1 (rows a–d) shows our updating and refinement of the HCGOM alignments of Patterson et al. (2006), by incorporating recent high-quality assembly sequence for C and M and by trimming 200 bp at the error-prone ends of the alignments. Improvement of data quality was indicated by reduction in pairwise distance estimates, better conformity with the molecular clock, and approximately equal estimates of the transition/transversion rate ratio κ (α/β in the notation of Kimura 1980) for all species pairs. It was apparent that sequencing errors show a lower κ than mutations. Our curated HC distance estimate (0.0121) was almost identical to the genome-wide estimate provided by the Chimpanzee Sequencing and Analysis Consortium (0.0123) (Mikkelsen et al. 2005).

This curation process yielded the inclusive autosomal data set *Complete*. From it, we selected our principal autosomal data set *Neutral* by the conventional method of masking functional regions. (For the X chromosome, a separate *Neutral X* data set was compiled by the same method.) Our second autosomal data set, *Inert TE*, contains only TEs that are well characterized and known to have been inert evolutionary fossils for at least 40 Myr. We also analyzed 4 partitions of *Inert TE* corresponding to the 4 major classes of primate TE: LINEs, SINEs, LTR retrotransposons, and DNA transposons.

Analysis of the *Neutral* Data Set

Basic Model

Both the basic model of Rannala and Yang (2003) and the extended models implemented in this article were applied to the *Neutral* data set, to examine the impact of model assumptions about neutrality, mutation rate variation, and recombination and to assess the sensitivity of posterior estimates to the prior. We describe these results first, which also serve to identify the best model for our data. We then present and discuss parameter estimates obtained under the best model from the *Complete*, *Inert TE*, and *Neutral X* data sets.

Under the basic model (Rannala and Yang 2003), we assigned the same diffuse gamma prior $G(2, 500)$ to the 4 θ parameters, with mean 0.0040 and 95% CI 0.0005–0.0111. Diffuse gamma priors were also assigned to the τ parameters: $G(4, 606)$ for τ_{HC}, $G(4, 465)$ for τ_{HCG}, $G(4, 219)$ for τ_{HCGO}, and $G(4, 131)$ for τ_{HCGOM}, with means 0.0066, 0.0086, 0.01826, and 0.0305, respectively. Those prior means were calculated using the species divergence times of Steiper and Young (2006) and the mutation rate 10^{-9} changes per site per year.

The posterior means of the parameters are shown in table 2, row a. The 95% equal-tail CIs are narrow around the mean, indicating the information content of this large data set. At the posterior mean parameter values, the gene tree for H, C, and G is expected to differ from the species tree at $P_{\mathrm{SG}} = 29\%$ of the loci (eq. 1). This is lower than the $P_{\mathrm{SE}} = 40\%$ in the phylogenetic analysis of a similar data set by Ebersberger et al. (2007), but as discussed earlier, P_{SE} is an overestimate of P_{SG} due to tree reconstruction errors. Ebersberger et al. then considered only loci at which the approximate posterior probability for the gene tree exceeded 0.95, obtaining the mismatch probability 0.23. This is a serious underestimate because conflicting gene trees tend to have shorter internal branch lengths and weaker support than matching gene trees, and application of the cutoff must have disproportionally removed conflicting gene trees. Our estimate, lying between 0.23 and 0.40, is thus qualitatively consistent with the estimates of Ebersberger et al.

Random Sequencing Errors or Violation of the Clock?

After data refinement, the *Neutral* data set still showed some branch length variation. By comparison with the H lineage and with reference to outgroup M, excess branch lengths were 0.0002 (=0.0627 – 0.0625) in C, 0.0013 in G, and 0.0023 in O (table 1). We analyzed the excess branch lengths by counting site patterns for species triplets (table 3). The proportions of sites with transitional (S) and transversional (V) differences were counted on each lineage. We subtracted the counts for H from the counts for each of C, G, and O to obtain excess proportions S and V for the C, G, and O lineages. Then κ was estimated under the K80 model

$$\hat{\kappa} = 2 \times \log(1 - 2S - V)/\log(1 - 2V) - 1 \qquad (6)$$

(Kimura 1980; Jukes 1987). This is close to $2S/V$ when S and V are small. The human sequence was assumed to have no errors, providing a good estimate of κ due to evolution, at 4.6. If all the excess in C was due to sequencing errors, κ for sequencing errors was ~1.6. Accordingly, the κ estimates suggest that the excess for O($\hat{\kappa}$=2.2) was mainly due to sequencing errors, whereas the excess for G($\hat{\kappa}$=3.6)

Table 2
Posterior Means and 95% CIs (in parentheses) of Parameters under Different Models for the *Neutral* Data Set

	θ_{HC}	θ_{HCG}	θ_{HCGO}	θ_{HCGOM}	τ_{HC}	τ_{HCG}	τ_{HCGO}	τ_{HCGOM}	ε_C	ε_G	ε_O	P_{SG}
(a) Basic model	6.5 (6.0–6.9)	3.4 (3.3–3.5)	6.1 (5.9–6.4)	11.8 (11.5–12.2)	4.1 (4.0–4.2)	6.7 (6.7–6.8)	14.6 (14.5–14.7)	26.4 (26.2–26.5)				0.29
(b) Sequencing errors	6.2 (5.8–6.6)	3.3 (3.2–3.4)	6.1 (5.8–6.3)	11.8 (11.5–12.2)	3.8 (3.7–3.9)	6.2 (6.1–6.3)	13.7 (13.6–13.8)	26.0 (25.8–26.1)	0.2 (0.1–0.3)	1.1 (1.0–1.2)	1.7 (1.5 –1.8)	0.31
(c) Fixed-rates model	6.3 (5.8–6.9)	3.5 (3.4–3.6)	6.1 (5.8–6.4)	NA	3.7 (3.5–3.8)	5.9 (5.8–6.0)	13.6 (13.5–13.8)	NA	0.6 (0.5–0.6)	1.6 (1.5–1.7)	Fixed	0.33
(d) Random-rates model	6.1 (5.7–6.6)	3.3 (3.2–3.4)	4.9 (4.7–5.2)	5.3 (4.8–5.8)	3.9 (3.8–4.0)	6.3 (6.2–6.4)	14.3 (14.2–14.5)	29.2 (29.0–29.5)	0.2 (0.1–0.3)	1.1 (1.0–1.2)	1.7 (1.6–1.9)	0.31
(e) Hominoid slowdown	6.1 (5.7–6.6)	3.2 (3.2–3.3)	4.7 (4.4–5.0)	6.8 (6.4–7.2)	3.9 (3.8–4.0)	6.3 (6.2–6.4)	14.5 (14.3–14.6)	24.4 (24.2–24.6)	0.2 (0.1–0.3)	1.1 (1.0–1.2)	1.7 (1.5–1.8)	0.31

NOTE.—NA, not applicable; θ, τ, and ε parameters are scaled by 10^3. In (e) Hominoid slowdown, $\varepsilon_M = 0.0082$ was fixed, whereas in other models, ε_M was not used. P_{SG} is the species tree–gene tree mismatch probability for H, C, and G, calculated using the parameter estimates by equation (1).

was more consistent with a higher evolutionary rate. This argument relied on the assumption that evolution and sequencing errors have distinct κ ratios. The eventual release of the finished gorilla and orangutan genome assemblies may reveal the true extent of clock violation in these species.

Parameters ε_C, ε_G, and ε_O were introduced to accommodate branch length variation for C, G, and O due to either species-specific sequencing errors or violation of the clock (fig. 1). A gamma prior $G(1, 1000)$, with mean 0.0010 and 95% CI 0.0003–0.0037, was assigned on each ε. The posterior means were 0.0002 for ε_C, 0.0011 for ε_G, and 0.0017 for ε_O (table 2). These were similar to the excess branch lengths obtained by simple distance calculations (0.0002, 0.0013, and 0.0023, respectively; see above and table 1).

The introduction of ε parameters had minimal effect on estimates of θs. The divergence times (τs) became smaller as excess branch lengths were accommodated by εs (table 2). As expected, εs and the corresponding τs had strong negative correlations (table 4). Furthermore, the priors on εs had virtually no effect on posterior estimates of θs and τs (supplementary table S1, Supplementary Material online). Overall, the effects of accommodating sequencing errors in the model were minor, partly because of the low error rates in the curated data.

The model of sequencing errors was also applied to the uncurated data set, statistics for which were shown in table 1, row a. This data set was known to include many sequencing errors as well as spurious alignments (see Methods). The results were shown in table 5. Under the basic model, θ and τ estimates differed considerably from estimates obtained from the curated *Neutral* data set. However, for all except the most distant ancestral population, the differences were greatly reduced under the model of sequencing errors (compare with table 2). As expected, the posterior estimates of error rates (εs) were much higher than for *Neutral*. Because the model performed well on this uncurated data set, it seemed likely that the model would successfully accommodate any remaining lower level of sequencing errors in the curated data set and that the inference was robust to a small number of spurious alignments that may have escaped the curation process.

Contribution of Ancestral Polymorphism to H–C Divergence

Note that the average coalescent time between 2 alleles is $2N$ generations, equivalent to a distance of $\theta/2(= 2Ng\mu)$. Therefore, the average H–C divergence δ_{HC} expected under the model is given by

$$\begin{aligned}\tfrac{1}{2}\delta_{HC} = {} & \tau_{HC} + \tfrac{1}{2}\theta_{HC} + \tfrac{1}{2}P_{HC}\left(\theta_{HCG} - \theta_{HC}\right)\\ & + \tfrac{1}{2}P_{HC}P_{HCG}\left(\theta_{HCGO} - \theta_{HCG}\right)\\ & + \tfrac{1}{2}P_{HC}P_{HCG}P_{HCGO}\left(\theta_{HCGOM} - \theta_{HCGO}\right)\\ = {} & \tau_{HC} + \tfrac{1}{2}\left(1 - P_{HC}\right)\theta_{HC} + \tfrac{1}{2}P_{HC}\left(1 - P_{HCG}\right)\theta_{HCG}\\ & + \tfrac{1}{2}P_{HC}P_{HCG}\left(1 - P_{HCGO}\right)\theta_{HCGO}\\ & + \tfrac{1}{2}P_{HC}P_{HCG}P_{HCGO}\theta_{HCGOM},\end{aligned} \tag{7}$$

where $P_{HC} = e^{-2(\tau_{HCG} - \tau_{HC})/\theta_{HC}}$ is the probability that 2 alleles entering the HC ancestral population do not coalesce in that

Table 3
Proportions of Sites with Transitional (*S*) and Transversional (*V*) Differences when the Human Is Compared with Another Ape in the *Neutral* Data Set

		Human Lineage			Excess in the Other Ape Lineage over Human		
Species Triplet	Total Sites	*S* (SXX)	*V* (VXX)	κ	*S* (XSX – SXX)	*V* (XVX – VXX)	κ
HC–M	23,285,394	0.003933	0.001727	4.57	0.000068	0.000084	1.61
HG–M	23,264,472	0.004931	0.002164	4.58	0.000778	0.000429	3.63
HO–M	14,034,931	0.010072	0.004363	4.66	0.001287	0.001152	2.24

NOTE.—H is compared with C, G, or O, with the outgroup M used as reference. For example, H and C are compared in the first row. The proportion of sites with pattern SXX (where human has a transitional difference while C and M are identical) is 0.003933. The proportion of sites with pattern VXX (where human has a transversional difference) is 0.001727. These 2 proportions give κ = 4.57 by equation (6). Similarly, transitions and transversions are counted for the C lineage. The excess proportions of transitions and transversions for the C lineage over the H lineage are shown in the table; these give the κ estimate 1.61. Rows 2 and 3 are similar comparisons for HG and HO.

population (see eq. 1), and $P_{\mathrm{HCG}}=\mathrm{e}^{-2(\tau_{\mathrm{HCGO}}-\tau_{\mathrm{HCG}})/\theta_{\mathrm{HCG}}}$ and $P_{\mathrm{HCGO}}=\mathrm{e}^{-2(\tau_{\mathrm{HCGOM}}-\tau_{\mathrm{HCGO}})/\theta_{\mathrm{HCGO}}}$ are defined similarly. With the parameter estimates from the *Neutral* data set under the basic model (table 2, row a), equation (7) gave $\frac{1}{2}\hat{\delta}_{\mathrm{HC}}=0.00410+0.00181+0.00074+0.00001+0.00000=0.00666$, implying that 39% of the divergence between H and C is due to ancestral polymorphism (contributed by the last 4 terms in eq. 7).

However, the predicted divergence $\frac{1}{2}\hat{\delta}_{\mathrm{HC}}=0.00666$ differed from the average JC69 distance $d_{\mathrm{HC}}/2$ = 0.01270/2 = 0.00635 of table 1. Here we treat the JC69 distances as observations to assess the fit of the model to data. Similar comparisons of predicted and observed H–G, H–O, and H–M distances did not show such a discrepancy (results not shown). We note that only the H–C and HC–G divergences are sufficiently close in time to generate a significant proportion of gene tree–species tree mismatches, and the tree mismatch information in the data is highly informative about the parameters in equation (1) (θ_{HC}, τ_{HC}, and τ_{HCG}). Under the basic model, it appeared that the excess branch length in G due to sequencing errors could not be reconciled with the tree mismatch information in the data, resulting in a poor fit to the observed HC distance.

The fit improved considerably when sequencing errors were accommodated in the model. As no errors were assumed on the H lineage, the predicted average H–C divergence was equal to the 5 terms of equation (7) plus $\frac{1}{2}\varepsilon_{\mathrm{C}}$, or $\frac{1}{2}\hat{\delta}_{\mathrm{HC}}=0.00385+0.00165+0.00077+0.00002+0.00000+0.00020/2=0.00639$ (using parameter estimates from table 2, row b). This was very close to the observed value $d_{\mathrm{HC}}/2$ = 0.00635 (table 1). The estimated contribution of ancestral polymorphism to H–C sequence divergence was unchanged, at 39%.

Rate Variation among Loci

We first implemented the fixed-rates model, using macaque to estimate relative rates (eq. 2) and analyzing data from 4 species only (HCGO). Again, sequencing errors in C and G were accommodated by ε_{C} and ε_{G}. With the removal of macaque, ε_{O} could not be estimated because it was confounded with τ_{HCGO}. Instead, ε_{O} was fixed at 0.00173, the estimate under the random-rates model (see below). The estimates of θ and τ for the 3 ancestral populations HC, HCG, and HCGO (table 2, row c) were similar to those obtained under the assumption of the same mutation rate for all loci (table 2, row b).

The fixed-rates model estimated relative rates using macaque data by ignoring polymorphism in the ancestor HCGOM, which might be substantial if N_{HCGOM} were large. We thus implemented the random-rates model, in which rates for loci were assigned the Dirichlet prior (eq. 4), with parameter α inversely related to the extent of rate variation among loci. We used α = 25, so that the relative rates have SD 0.2 in the prior, slightly lower than the empirical SD (0.278) calculated by equation (2) for the fixed-rates model. The estimate for θ_{HCGOM} (0.0053) was about half that obtained under the basic model. The inversely correlated parameter τ_{HCGOM} increased slightly from 0.0260 to 0.0292. Parameters for more recent ancestral populations (HC, HCG, HCGO) were well estimated and changed little (table 2, row d).

Table 4
Correlation between Parameters in the Posterior Distribution

	θ_{HC}	θ_{HCG}	θ_{HCGO}	θ_{HCGOM}	τ_{HC}	τ_{HCG}	τ_{HCGO}	τ_{HCGOM}	ε_{C}	ε_{G}
θ_{HCG}	−0.26									
θ_{HCGO}	0.01	−0.04								
θ_{HCGOM}	0.00	−0.03	−0.12							
τ_{HC}	**−0.77**	0.19	−0.02	0.03						
τ_{HCG}	0.20	**−0.42**	0.00	0.06	0.21					
τ_{HCGO}	0.01	0.12	**−0.67**	0.10	0.18	0.31				
τ_{HCGOM}	0.00	0.05	0.13	**−0.90**	0.03	0.05	0.01			
ε_{C}	0.01	0.00	0.02	−0.05	**−0.41**	**−0.53**	−0.27	−0.03		
ε_{G}	−0.03	0.00	0.02	−0.05	−0.25	**−0.59**	−0.28	−0.04	0.40	
ε_{O}	−0.01	−0.08	−0.02	−0.03	−0.15	−0.27	**−0.43**	−0.09	0.23	0.24

NOTE.—The *Neutral* data set was analyzed under the random-rates model (Table 2, row d). Strong correlations are highlighted in bold.

Table 5
Posterior Means and 95% CIs (in parentheses) for Parameters Obtained from the Uncurated Data

	θ_{HC}	θ_{HCG}	θ_{HCGO}	θ_{HCGOM}	τ_{HC}	τ_{HCG}	τ_{HCGO}	τ_{HCGOM}	ε_C	ε_G	ε_O	P_{SG}
Basic model	7.0 (6.8–7.2)	4.4 (4.4–4.5)	7.2 (7.1–7.)	14.8 (14.6–15.0)	4.6 (4.6–4.7)	7.8 (7.8–7.8)	15.6 (15.5–15.7)	26.6 (26.5–26.7)				0.27
Sequencing errors	6.2 (6.0–6.4)	3.8 (3.8–3.9)	5.6 (5.5–5.8)	8.7 (8.4–9.0)	3.8 (3.8–3.9)	6.2 (6.2–6.3)	14.3 (14.3–14.4)	28.8 (28.7–29.0)	1.1 (1.0–1.1)	3.9 (3.8–3.9)	2.8 (2.7–2.9)	0.31

Note.—θ, τ, and ε parameters are scaled by 10^3

We changed α in the Dirichlet prior to assess the sensitivity to the prior (supplementary table S2, Supplementary Material online). Estimates of θ_{HCGOM}, and to a lesser degree τ_{HCGOM}, were sensitive to α. However, parameter estimates for the other populations (HC, HCG, HCGO) were stable, showing little change with wide variation in the prior. We concluded that this data set contained sufficient information on rate variation among loci to obtain reliable estimates for all populations except the earliest, HCGOM. Accordingly, estimates for θ_{HCGOM} were considered unreliable and were discarded from our final results; τ_{HCGOM} estimates were retained but should be considered approximate.

Hominoid Slowdown

There is considerable evidence for a rate difference between OWMs and the great apes (Goodman 1961; Li et al. 1996; Yi et al. 2002; Kim et al. 2006). Our data set lacked an appropriate outgroup to provide new information about such a rate difference; but if it exists, it might affect our estimates. We ran MCMCCOAL with $\varepsilon_M = 0.0082$ fixed. This represented a one-third higher mutation rate on the macaque lineage than in the apes. As would be expected, the posterior estimate of τ_{HCGOM} became much smaller (table 2, row e). However, our estimates for population sizes and speciation times in the 3 more recent ancestral populations (HC, HCG, HCGO) were robust to the proposed large difference in mutation rate between OWM and the hominoid lineage.

Robustness to Priors

The effect of θ and τ priors on posterior parameter estimates was assessed in supplementary tables S3 and S4, Supplementary Material online. Again, posterior distributions were found to be robust to drastic changes in the priors. In summary, we found robustness to all priors and narrow confidence intervals, reflecting the large amount of information in the data. We concluded that accommodating both excess branch lengths and rate variation among loci did not overspecify the model for the *Neutral* data set.

Locus Length and Within-Locus Recombination

With an average locus length of 508 bp (*Neutral* data set), the assumption of no recombination within a locus is unrealistic over the timescale of hominoid evolution. In order to assess the impact of recombination, we generated data sets with progressively shorter loci by sampling one segment of a specified length from a random position within each locus in the *Neutral* data set. We predicted that if recombination were a serious problem, the parameter estimates would vary considerably with locus length because recombination would have a greater impact on longer segments. The results of this analysis are shown in table 6. The 95% CIs became wider when the segment size was reduced, reflecting the reduced information content in the data. Nevertheless, the posterior means for θs were not very different from those obtained from the full-length data set. Thus, our parameter estimates did not appear to be greatly

Table 6
Posterior Means and 95% CIs (in parentheses) of Parameters when Shorter Loci Were Sampled from the *Neutral* Data Set

Locus Length	θ_{HC}	θ_{HCG}	θ_{HCGO}	τ_{HC}	τ_{HCG}	τ_{HCGO}	τ_{HCGOM}	P_{SG}
508	6.2 (5.8–6.6)	3.3 (3.2–3.4)	6.1 (5.8–6.3)	3.8 (3.7–3.9)	6.2 (6.1–6.3)	13.7 (13.6–13.8)	26.0 (25.8–26.1)	0.31
333	6.7 (6.1–7.4)	3.5 (3.3–3.6)	5.9 (5.6–6.2)	3.7 (3.5–3.8)	6.0 (6.0–6.1)	13.6 (13.4–13.8)	25.5 (25.3–25.7)	0.33
200	7.3 (6.4–8.3)	3.7 (3.5–3.8)	6.5 (6.2–6.9)	3.5 (3.3–3.7)	5.9 (5.8–6.0)	13.2 (13.0–13.4)	25.2 (24.9–25.4)	0.35
100	7.8 (6.1–9.6)	4.2 (4.0–4.4)	7.1 (6.6–7.6)	3.4 (3.1–3.7)	5.6 (5.5–5.8)	12.9 (12.6–13.2)	24.5 (24.2–24.9)	0.37
50	6.9 (4.4–10.1)	4.6 (4.2–5.0)	8.3 (7.5–9.1)	3.4 (2.9–3.9)	5.4 (5.2–5.6)	12.2 (11.7–12.6)	23.7 (23.1–24.2)	0.38

NOTE.—The model of sequencing errors is used, with θ and τ parameters scaled by 10^3 and results for θ_{HCGOM} are not shown. For the samples, error rates are not estimated but are fixed at the posterior means estimated for the full-length data set (table 2, second row). P_{SG} is the tree mismatch probability for H, C, and G.

sensitive to the effects of within-locus recombination. However, ignoring model violations such as within-locus recombination may have caused the 95% CIs in our main analyses to be too narrow.

This analysis also highlighted the superiority of the full likelihood method over the tree-mismatch method (see Introduction). Our Bayesian analysis produced quite reasonable estimates even when each locus contained only 50 sites, with on average less than one pairwise difference between H, C, and G. The probability P_{SG} of mismatch between the species tree and the gene tree, calculated from the parameter estimates, increased only slightly from 31% for the full-length data to 38% for the shortest loci (table 6 and fig. 2). However, the probability P_{SE} of mismatch between the species tree and the estimated gene tree, which is used by the tree-mismatch method, increased to 59% (fig. 2). It was noteworthy that even with the full-length data, P_{SE} (0.41) was considerably higher than P_{SG} (0.31), so errors in tree reconstruction have a considerable impact on the simple tree-mismatch method.

Analysis of the *Complete* and *Inert TE* Data Sets

From the above analysis, we consider the random-rates model incorporating sequencing errors to be our best model. This was used to analyze the *Complete* and *Inert TE* data sets, with results shown in table 7. For *Complete*, estimates of θs and τs were all slightly smaller than those from the *Neutral* data set. This can be explained by the presence of a small proportion of masked sites under purifying selection. The ratios of τ estimates between the 2 data sets suggested that the mutation rate of *Complete* was about 0.97 times that of *Neutral*. Aside from this rate difference, the parameter estimates were almost identical between the 2 data sets. The estimates were evidently robust to any violation of the neutral assumption in this large data set.

Estimates of θs and τs from the *Inert TE* data set are comparable with those from the *Neutral* and *Complete* data sets (table 7). However, the estimates suggest some variation in mutation rate among the 4 classes of TEs. We discuss the results below using a calibration to the H–C speciation time.

Speciation Times and Ancestral Population Sizes

To translate θs and τs into population size *N*s and divergence time *T*s, it is necessary to use a mutation rate μ or a divergence time for calibration. We calibrated to the human–chimpanzee divergence time (T_{HC}), with 2 values used: 4 and 6 Myr. Four million years is the minimum time consistent with widely accepted fossil evidence for H–C divergence (Leakey et al. 1995), whereas more controversial fossil evidence (Senut et al. 2001; Brunet et al. 2002) might require an earlier divergence at about 6 Myr. The generation time was assumed to be $g = 15$ years. Modern humans have longer generation times, but ancestral species were probably physically smaller, perhaps similar to modern macaques, which have $g \approx 11$ years (Gage 1998). The calibrated results are shown in table 8.

Calibration to $T_{HC} = 4$ Myr implied a mutation rate of 0.98×10^{-9} per site per year for the *Neutral* data set. This was similar to the commonly used rate $\mu = 10^{-9}$ per site per year (e.g., Takahata et al. 1995; Rannala and Yang 2003). The population size for the HC ancestor (N_{HC}) was estimated to be $\sim 10^5$, about 10 times larger than estimates for modern humans. This is consistent with Takahata's (1993) early estimate from analysis of MHC alleles and is much larger than the estimates of Yang (2002) and Rannala and Yang (2003) from the data of 53 loci of Chen and Li (2001). The latter data set may be atypical in having an

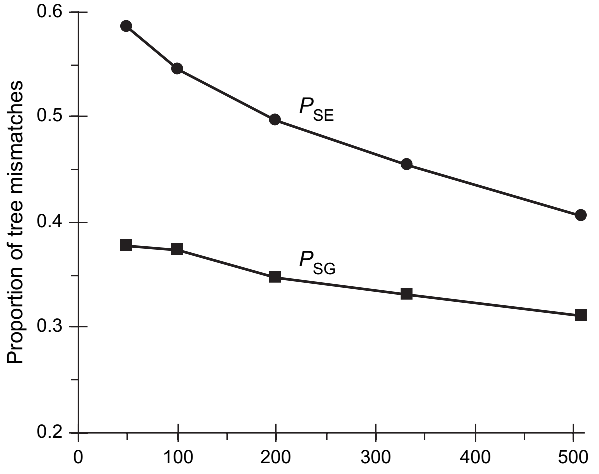

FIG. 2.—Errors in phylogenetic tree reconstruction inflate the species tree–gene tree mismatch probability. P_{SG} is the probability of mismatch between the species tree and the gene tree, derived from the coalescent parameter estimates by equation (1), whereas P_{SE} is the probability of mismatch between the species tree and the estimated gene tree. Both refer to the phylogenetic relationship among H, C, and G. Data sets with progressively shorter loci, used in table 6, are analyzed.

Table 7
Posterior Means and 95% CIs (in parentheses) for Parameters Obtained from Different Data Sets

	θ_{HC}	θ_{HCG}	θ_{HCGO}	τ_{HC}	τ_{HCG}	τ_{HCGO}	τ_{HCGOM}	ε_C	ε_G	ε_O	P_{SG}
Complete	5.6 (5.4–5.9)	3.2 (3.1–3.2)	4.8 (4.7–5.0)	3.8 (3.8–3.9)	6.1 (6.1–6.2)	13.9 (13.9–14.0)	28.2 (28.1–28.4)	0.1 (0.1–0.2)	1.1 (1.0–1.1)	1.5 (1.4–1.6)	0.29
Neutral	**6.1 (5.7–6.6)**	**3.3 (3.2–3.4)**	**4.9 (4.7–5.2)**	**3.9 (3.8–4.0)**	**6.3 (6.2–6.4)**	**14.3 (14.2–14.5)**	**29.2 (29.0–29.5)**	**0.2 (0.1–0.3)**	**1.1 (1.0–1.2)**	**1.7 (1.6–1.9)**	**0.31**
Inert TE	**6.8 (6.3–7.3)**	**3.7 (3.6–3.8)**	**6.1 (5.8–6.4)**	**3.7 (3.6–3.8)**	6.2 (6.1–6.3)	**14.4 (14.2–14.6)**	**30.9 (30.6–31.2)**	**0.3 (0.2–0.4)**	**1.2 (1.1–1.3)**	**1.7 (1.5–1.9)**	**0.32**
LINE	6.6 (5.9–7.3)	3.3 (3.2–3.5)	5.6 (5.2–6.0)	3.4 (3.2–3.5)	5.7 (5.7–5.8)	13.1 (12.9–13.3)	29.6 (29.0–30.0)				
SINE	7.5 (6.6–8.5)	4.5 (4.3–4.7)	7.1 (6.6–7.7)	4.0 (3.8–4.2)	6.7 (6.6–6.8)	16.3 (16.0–16.6)	33.8 (33.3–34.2)				
LTR	7.3 (6.1–8.6)	3.7 (3.5–4.0)	5.9 (5.3–6.6)	3.8 (3.5–4.0)	6.3 (6.1–6.4)	14.3 (14.0–14.7)	32.9 (31.8–33.3)				
DNA	6.1 (4.6–7.7)	3.6 (3.3–4.0)	6.4 (5.5–7.4)	3.6 (3.3–3.9)	5.9 (5.7–6.0)	13.2 (12.7–13.6)	30.9 (29.9–31.6)				
Neutral X	**2.6 (2.0–3.3)**	**2.0 (1.6, 2.5)**	**3.2 (1.9, 4.4)**	**3.1 (2.8–3.9)**	**5.4 (5.1–5.6)**	**11.7 (11.1–12.3)**	**25.5 (24.3–26.6)**				

NOTE.—The model of sequencing errors is used, with θ, τ, and ε parameters scaled by 10^3 and results for θ_{HCGOM} are not shown. Parameters ε_C, ε_G, and ε_O for TE subclasses were not estimated but were fixed at the posterior means for the entire *Inert TE* data set. Those for *Neutral X* were fixed at the posterior means for the *Neutral* data set. P_{SG} is the tree mismatch probability for H, C, and G.

unusually small variance across loci in sequence divergences and a small gene tree–species tree mismatch proportion (P_{SE} = 0.203) (Yang 2002; Satta et al. 2004). Estimates for earlier ancestral populations N_{HCG} (55,000) and N_{HCGO} (85,000) were somewhat smaller. The species divergence times were estimated to be 6.4 Myr for gorilla, 14.6 Myr for orangutan, and ~25–30 Myr for macaque.

When calibrated to T_{HC}, estimates from the *Complete* data set were very similar to the *Neutral* data set. Application of the same calibration to the 4 classes of TEs suggests that they have different mutation rates (table 8). LINE elements have the lowest mutation rate, with the rate for DNA transposons 6% higher, LTR transposons 11% higher, and SINEs 17% higher. When this rate difference is accounted for, the 4 TE classes gave very similar estimates of *N*s and *T*s, which are somewhat higher than the estimates from the *Neutral* data set (by about 25% for *N*s and less for *T*s). One reason could be the smaller mean locus length (321 vs. 508 bp), so that the modeling error due to within-locus recombination is reduced (see table 6 for the impact of locus length).

Calibrating to T_{HC} = 6 Myr instead, the mutation rate for the *Neutral* data set was only 2/3 as large, at 0.65×10^{-9} mutations per site per year, implying that all population sizes and divergence times were 1.5 times as large as under the calibration T_{HC} = 4 Myr. The divergence time of 45 Myr for macaque (T_{HCGOM}) seemed too old, although the hominoid slowdown model with a one-third higher rate for macaque reduced the estimate to 37 Myr.

The calibration T_{HC} = 4 Myr may be preferable, as seemingly incompatible fossil evidence (Senut et al. 2001; Brunet et al. 2002) may be due to lineage sorting in hominoid ancestors. Ebersberger et al. (2007) discussed the intriguing impact of widespread ancestral polymorphism and lineage sorting on the interpretation of morphological evolution and fossil data. The authors noted that coding and regulatory regions of the genome were affected by lineage sorting as much as neutral regions even though the former had lower rates of evolution. Some morphological traits specific to modern humans arose by genetic mutation in the HC common ancestor, with polymorphism maintained through the time of H–C speciation. Genetic and corresponding morphological polymorphisms could then be present in both species until eventually one allele and the corresponding trait were fixed in each species. Thus, early fossil remains, despite the presence of apomorphies specific to modern humans, might be from either the human or chimpanzee lineage or from their common ancestor.

X Chromosome Divergence and Human–Chimpanzee Speciation

We first considered the ratio of mean genetic divergence between X and the autosomes (A), $d_{(X)}/d_{(A)}$, calculated directly from JC69 distance estimates in species comparisons (table 1 last row). The $d_{(X)}/d_{(A)}$ ratio was 0.70 for H–C, 0.80–0.82 between any other 2 apes, and 0.84–0.85 between any ape and macaque. Under a coalescent model with one ancestral population size, the $d_{(X)}/d_{(A)}$

Table 8
Posterior Means and 95% CIs (in parentheses) for Hominoid Effective Population Sizes N (×1,000) and Speciation Times T (Myr) using 2 calibrations: H–C speciation at 4 or 6 Myr

Data Set	μ	N_{HC}	N_{HCG}	N_{HCGO}	T_{HCG}	T_{HCGO}	T_{HCGOM}
Calibration: $T_{HC} = 4$ Myr							
Complete	0.95	99 (95–102)	55 (54–56)	85 (82–87)	6.4 (6.4–6.5)	14.6 (14.5–14.7)	29.6 (29.4–29.8)
Neutral	**0.98**	**104 (97–112)**	**55 (54–57)**	**84 (80–89)**	**6.4 (6.4–6.5)**	**14.6 (14.5–14.8)**	**29.9 (29.6–30.1)**
With hominoid slowdown	0.98	104 (97–111)	55 (54–57)	79 (75–84)	6.4 (6.3–6.5)	14.7 (14.6–14.9)	24.8 (24.6–25.0)
Inert TE	**0.93**	**122 (113–131)**	**66 (64–68)**	**110 (104–115)**	**6.7 (6.6–6.8)**	**15.6 (15.4–15.8)**	**33.3 (33.0–33.6)**
LINE	0.85	129 (115–144)	66 (63–69)	110 (102–118)	6.8 (6.7–6.9)	15.4 (15.2–15.7)	34.8 (34.2–35.4)
SINE	0.99	127 (111–143)	76 (72–79)	120 (110–129)	6.8 (6.7–6.9)	16.4 (16.1–16.7)	34.0 (33.5–34.5)
LTR	0.94	128 (107–152)	65 (61–70)	105 (93–117)	6.7 (6.5–6.8)	15.2 (14.8–15.6)	34.9 (33.7–35.4)
DNA	0.90	112 (86–142)	67 (61–73)	119 (102–136)	6.5 (6.3–6.7)	14.6 (14.1–15.1)	34.4 (33.3–35.1)
Calibration: $T_{HC} = 6$ Myr							
Complete	0.64	148 (142–154)	83 (81–84)	127 (123–131)	9.6 (9.6–9.7)	21.9 (21.8–22.1)	44.4 (44.1–44.6)
Neutral	**0.65**	**157 (146–168)**	**83 (81–86)**	**126 (120–133)**	**9.6 (9.5–9.7)**	**22.0 (21.7–22.2)**	**44.8 (44.4–45.2)**
With hominoid slowdown	0.65	156 (145–167)	83 (80–85)	119 (112–127)	9.6 (9.5–9.7)	22.1 (21.9–22.4)	37.2 (36.9–37.5)
Inert TE	**0.62**	**182 (169–196)**	**99 (96—102)**	**164 (156–173)**	**10.1 (9.9–10.2)**	**23.4 (23.1–23.6)**	**49.9 (49.5–50.4)**
LINE	0.57	194 (173–216)	99 (94–103)	165 (152–177)	10.1 (10.0–10.3)	23.2 (22.8–23.5)	52.3 (51.3–53.0)
SINE	0.66	190 (167–215)	113 (108–119)	180 (165–194)	10.2 (10.0–10.3)	24.6 (24.1–25.0)	51.0 (50.3–51.7)
LTR	0.63	192 (161–228)	98 (92–105)	157 (139–175)	10.0 (9.8–10.2)	22.8 (22.3–23.3)	52.3 (50.6–53.1)
DNA	0.60	168 (129–213)	100 (91–110)	179 (154–204)	9.8 (9.5–10.1)	21.9 (21.2–22.7)	51.5 (49.9–52.7)

NOTE.—The *random-rates* model is assumed. Average rate among loci μ is scaled by 10^{-9} mutations per site per year.

ratio may be expressed in terms of the population parameters as

$$\frac{d_{(X)}}{d_{(A)}} = \frac{\tau_{(X)} + \frac{1}{2}\theta_{(X)}}{\tau_{(A)} + \frac{1}{2}\theta_{(A)}} = \frac{\mu_{(X)}}{\mu_{(A)}} \times \frac{T_{(X)} + 2N_{(X)}}{T_{(A)} + 2N_{(A)}}. \qquad (8)$$

(A more accurate formula may be constructed using eq. 7 to account for several ancestral population sizes.) Therefore, $d_{(X)}/d_{(A)}$ is affected by the mutation rate ratio $\mu_{(X)}/\mu_{(A)}$, the ratio of speciation times $T_{(X)}/T_{(A)}$, and the ratio of effective population sizes $N_{(X)}/N_{(A)}$. Our model allowed estimation from the data of 4 τ and 4 θ parameters for the 4 ancestral populations separately for X and for A, without imposing extrinsic assumptions about parameter values.

We applied the random-rates model to the *Neutral X* data set, using the same priors on θs and τs as for the autosomal analyses. (Although systematic differences were expected between the X and the autosomes, the priors were diffuse and the posterior was insensitive to wide variation in the prior mean.) Parameters ε_C, ε_G, and ε_O representing sequencing errors were not estimated but were fixed at the posterior means obtained for the *Neutral* data set. The results are shown in table 7, last row. Estimates of θs and τs for the X were lower than for the autosomes (A).

To assess the relative importance of various factors in causing the small $d_{(X)}/d_{(A)}$ ratio for H–C, we first calculated the ratios $\tau_{(X)}/\tau_{(A)}$ for the 4 ancestral species: 0.79 for HC, 0.86 for HCG, 0.81 for HCGO, and 0.87 for HCGOM (table 9). These ratios were similar, and the average (0.83) lay within the 95% CIs for all 4 populations. In theory, a formal hypothesis testing or Bayesian model comparison can be used to compare 2 nested models. The more general model is the separate analysis of the *Neutral* and *Neutral X* data, estimating a set of 4 τ and 4 θ parameters for each. The null model assumes that all 4 speciation times are the same between X and A and that $\mu_{(X)}/\mu_{(A)}$ is constant. This null model is equivalent to placing the constraint on the parameters of the general model that $\tau_{(X)}/\tau_{(A)}$ is constant across all 4 ancestors. This is a model of simple speciation (see Discussion). Although the general model is implemented to produce results of table 7, the null model has not been implemented. Instead, we use a less rigorous argument: the fact that the estimated $\tau_{(X)}/\tau_{(A)}$ ratios are similar and their mean is inside all the 95% CIs suggests that a Bayesian model comparison would favor the simpler model of lower dimension (Dawid 1999). Our estimates are therefore consistent with the simple null model of simultaneous speciation. A similar conclusion was reached by Innan and Watanabe (2006) who applied a model of gene flow to neutral data from human and chimpanzee.

Because $T_{(X)} = T_{(A)}$, our estimates of $\tau_{(X)}/\tau_{(A)}$ were estimates of the X/A mutation rate ratio $\mu_{(X)}/\mu_{(A)}$. The mean

Table 9
Posterior Means and 95% CIs (in parentheses) for Parameter Ratios between X Chromosome and Autosomes

Population	H	HC	HCG	HCGO	HCGOM
$\tau_{(X)}/\tau_{(A)}$	NA	0.79 (0.72–0.86)	0.86 (0.82–0.90)	0.81 (0.77–0.86)	0.87 (0.83–0.91)
$\theta_{(X)}/\theta_{(A)}$	0.61	0.43 (0.33–0.54)	0.62 (0.50–0.76)	0.64 (0.39–0.89)	Not shown
$N_{(X)}/N_{(A)}$	0.73	0.51 (0.39–0.65)	0.75 (0.60–0.91)	0.77 (0.47–1.06)	

NOTE.—NA, not applicable. The *Neutral X* and *Neutral* data sets were analyzed under the random-rates model. The posterior distribution for $\tau_{(X)}/\tau_{(A)}$ (and similarly for $\theta_{(X)}/\theta_{(A)}$) was constructed by independent random sampling from the posterior distributions of $\tau_{(X)}$ and $\tau_{(A)}$, generated in the MCMC. This gave almost identical results to the approach of diving the 2.5 and 97.5 percentiles of $\tau_{(X)}$ by the posterior mean of $\tau_{(A)}$, possibly because the posterior distributions of both $\tau_{(X)}$ and $\tau_{(A)}$ are highly concentrated. The mean of $\tau_{(X)}/\tau_{(A)}$ across the 4 populations, 0.83, was used to estimate $\mu_{(X)}/\mu_{(A)}$, which was then used to convert $\theta_{(X)}/\theta_{(A)}$ into $N_{(X)}/N_{(A)}$ for each population. The estimate for $\theta_{(X)}/\theta_{(A)}$ in modern humans is from Sachidanandam et al. (2001).

estimate was $\mu_{(X)}/\mu_{(A)} = 0.83$. The mutation rates on X and on A are expected to be different, with $\mu_{(X)} < \mu_{(A)}$, because X chromosomes spend more time in females, experiencing a lower mutation rate due to male mutation bias (Haldane 1935; see Ellegren 2007 for review). Let α be the male/female mutation rate ratio. Then with X spending 2/3 of the time in females and 1/3 in males, whereas an autosome spends half of the time in each, it is expected that $\mu_{(X)}/\mu_{(A)} = \left(\frac{2}{3} + \frac{1}{3}\alpha\right)/\left(\frac{1}{2} + \frac{1}{2}\alpha\right)$ or

$$\alpha = \frac{4 - 3\mu_{(X)}/\mu_{(A)}}{3\mu_{(X)}/\mu_{(A)} - 2} \qquad (9)$$

(Miyata et al. 1987).

By this equation, our estimate $\mu_{(X)}/\mu_{(A)} = 0.83$ implied $\hat{\alpha} = 3.0$. This lay within the range 2–5 obtained in several previous studies (Lander et al. 2001; Goetting-Minesky and Makova 2006; Taylor et al. 2006). Our estimate was higher than the 1.9 calculated by Patterson et al. (2006: supplementary note 8), who used $\mu_{(X)}/\mu_{(A)} = 0.899$, their $d_{(X)}/d_{(A)}$ ratio for H–M. The $d_{(X)}/d_{(A)}$ ratio for H–M from our data was 0.85 (table 1, last row), which would imply $\hat{\alpha} = 2.6$ by the method of Patterson et al. In any case, this method ignores the effects of ancestral polymorphism and is expected to overestimate $\mu_{(X)}/\mu_{(A)}$ and underestimate α (see eq. 8, noting that $N_{(X)} < N_{(A)}$).

Note that $\mu_{(X)}/\mu_{(A)}$ has a narrow range from 4/3 to 2/3 corresponding to α from 0 to ∞, so α is rather sensitive to $\mu_{(X)}/\mu_{(A)}$ (eq. 9). The 95% CIs for $\tau_{(X)}/\tau_{(A)}$, while consistent with a constant $\mu_{(X)}/\mu_{(A)}$, cannot exclude some variation in $\mu_{(X)}/\mu_{(A)}$ between ancestors, which if present might correspond to considerable variation in α. This caveat is applicable to any estimate for α derived from a comparison between X and A. However, our results below showed that such variation in α, if present, was not the principal cause of the low $d_{(X)}/d_{(A)}$ seen for H–C.

We next considered the ratios $\theta_{(X)}/\theta_{(A)}$ for 3 ancestral populations. The posterior means and 95% CIs are shown in table 9. The ratios for HCG (0.62) and HCGO (0.64) were similar to the estimate (0.61) obtained for modern humans by Sachidanandam et al. (2001). We adjusted these ratios by our estimated mutation rate ratio $\mu_{(X)}/\mu_{(A)} = 0.83$ to obtain the ratio of effective population size, $N_{(X)}/N_{(A)}$. This was ≈3/4, as expected under a simple model with balanced sex ratio, in which the population contains 3 X chromosomes to every 4 autosomes. However, for the HC common ancestor, $\theta_{(X)}/\theta_{(A)}$ was considerably lower, at 0.43. Adjusting by $\mu_{(X)}/\mu_{(A)} = 0.83$ implied $N_{(X)}/N_{(A)} = 0.51$. Even if α were considerably higher than our estimate of 3.0, say $\alpha = 6.0$ (which would imply $\mu_{(X)}/\mu_{(A)} = 0.76$ by eq.9), the estimate for $N_{(X)}/N_{(A)}$ would only increase to 0.56, still much lower than the expected value of 3/4.

Our estimates imply an unusually small X chromosome effective population size in the HC common ancestor. Furthermore, the size reduction was transient in the HC population, not inferred in earlier ancestors HCG or HCGO or in modern humans (Sachidanandam et al. 2001). A similar result of low relative diversity for the X chromosome in HC and normal relative diversity in HCG was obtained by Hobolth et al. (2007), who implemented an approximate hidden Markov chain model to deal with recombination and applied it to a somewhat smaller data set.

Discussions

The Power of Joint Likelihood Analysis of Multiple Species

We made 2 significant improvements to the data set of Patterson et al. (2006): incorporation of recent assembly sequence and recuration of the data to obtain high-quality continuous alignments instead of single sites. Furthermore, we used a new method to compile a data set containing well-characterized inert elements rather than uncharacterized presumptively neutral sequence. The broad similarity of parameter estimates between the data sets suggested that the results were reliable.

We extended the Bayesian coalescent model of Rannala and Yang (2003) to accommodate variable mutation rates among loci and to account for sequencing errors or violation of the clock. Using the *Neutral* data set, we showed that posterior parameter estimates were insensitive to wide variation in the priors. An exception was the earliest ancestral population size, where outgroup information was absent. We considered the impact of model assumptions concerning recombination. We ensured adequate locus separation (10 kb) to satisfy the assumption of *free recombination between loci* and sampled short segments from each locus to verify the robustness of our results to potential violations in the assumption of *no recombination within a locus*. This analysis allowed us for the first time to obtain robust and precise population size estimates for the entire hominoid lineage extending back to the cercopithecoid divergence approximately 30 Myr ago. Throughout this period, hominoid ancestral populations were shown to be an order of magnitude larger than modern humans.

Our analysis demonstrated the importance of adopting a probabilistic model and a rigorous statistical methodology in inference problems that involve multiple parameters with complex dependence structures. Indeed the power of the full likelihood approach is manifest only in joint analysis of sequence data from multiple species. When only 2 species are analyzed, estimation of the ancestral θ may be critically compromised by the strong correlation between θ and τ, and by rate variation among loci, which is confounded with stochastic variation of coalescence times (Yang 1997a). Such sensitivities exist also in analysis of more than 2 species (table 4 and supplementary table S2, Supplementary Material online), but the effects are mostly limited to the population parameter and divergence time of the earliest ancestor (θ_{HCGOM} and τ_{HCGOM} in our example), whereas the parameters for more recent populations are barely affected.

We found it valuable to model sequencing errors explicitly in the analysis. When the error rate is comparable to the level of natural polymorphisms, population genetics inference may be seriously affected. We performed a heuristic analysis of the extent of sequencing errors in our data, by using the observation that the transition/transversion rate ratio due to evolution differs from that due to sequencing errors. In general, evolution and sequencing errors may have different characteristics and such differences may be accommodated in the model. Experimental data may now exist which would allow analysis of empirical patterns of sequencing errors for building better models. This issue may become even more important with the availability of new high-throughput sequencing technologies with higher error rates at low coverage, with characteristic error distributions that differ between technologies (Johnson and Slatkin 2008).

Human–Chimpanzee Speciation and Transient Reduction of Effective Population Size in the HC Ancestor

Patterson et al. (2006) suggested that the human–chimpanzee speciation process was complex, with extensive hybridization following initial separation of the 2 species. This provocative hypothesis was based on 2 observations: 1) the high variation in sequence divergence throughout the genome and, in particular, 2) extreme reduction in sequence divergence on X relative to the autosomes ($d_{(X)}/d_{(A)}$) for H–C. Wakeley (2008) pointed out that the null hypothesis of simple speciation was never tested or rejected by Patterson et al. (2006). As correctly observed by Barton (2006), the variation in sequence divergence across autosomal loci can be explained by stochastic fluctuation in the coalescent process in the common ancestors under a model of simple speciation without introgression. As our results have shown, the simple speciation model predicts considerable variation between loci, with the gene tree differing from the species tree ((HC)G) at about one-third of loci and with ancestral polymorphism accounting for 39% of the average H–C sequence divergence. We now turn to the more complex analysis of evidence for reduced divergence on X between human and chimpanzee.

The $d_{(X)}/d_{(A)}$ ratio for H–C calculated from table 1 of Patterson et al. (2006) was $0.0053/0.0070 = 0.750$. (Wakeley 2008 calculated this to be 0.76 for the same data, the slight difference being probably due to different corrections for recurrent mutations.) The ratio from our data was $0.0088/0.0127 = 0.70$ for H–C and 0.82 for H–G (table 1, last row). Note that even though the sequence distances in the data of Patterson et al. (2006) were much smaller than ours due to those authors' removal of hypermutable CpG sites, the $d_{(X)}/d_{(A)}$ ratios were similar.

By equation (8), the small $d_{(X)}/d_{(A)}$ ratio for H–C could be due to any of the following ratios being small: $\mu_{(X)}/\mu_{(A)}$, $T_{(X)}/T_{(A)}$, or $N_{(X)}/N_{(A)}$. The methods used by Patterson et al. (2006), which rely on counting variable (divergent) sites and calculating distances between species pairs, may not be powerful enough to disentangle the relative contributions of these factors. To account for the mutation rate difference between the X and the autosomes, Patterson et al. (2006: supplementary note 8) rescaled the $d_{(X)}/d_{(A)}$ ratios for H–C and H–G by the ratio for H–M, yielding 0.835 ($=0.750/0.899$) for H–C and 0.977 for H–G. The values from our data are similar, at 0.82 ($=0.70/0.85$) for H–C and 0.96 ($=0.82/0.85$) for H–G (see table 1, last row). This procedure is equivalent to treating the $d_{(X)}/d_{(A)}$ ratio for H–M (0.899) as a direct estimate for $\mu_{(X)}/\mu_{(A)}$, ignoring polymorphism in the ancestor HCGOM, to calculate the scaled ratio, $\frac{d_{(X)}}{d_{(A)}} \Big/ \frac{\mu_{(X)}}{\mu_{(A)}} = \frac{T_{(X)}+2N_{(X)}}{T_{(A)}+2N_{(A)}}$ in our notation. Patterson et al. (2006: supplementary note 2) estimated ranges of parameter values under a simple demographic model to assess whether it was compatible with the observed X/A ratios. The authors' use of only total counts of variable sites (instead of the original sequence alignments) meant that the model was unidentifiable, so that a number of constraints on the parameter values were applied. As the observed X/A ratio for H–G could be predicted by the model under reasonable parameter values, Patterson et al. (2006) considered it unlikely for the mutation rate to have changed during the evolution of the great apes and thus ruled out variation in $\mu_{(X)}/\mu_{(A)}$ (or variation in α) as a possible reason for the low $d_{(X)}/d_{(A)}$ ratio for H–C. This procedure does not constitute a statistical test, in which one would estimate the parameters from the data and use the estimates to generate the null distribution of the test statistic (see also Wakeley 2008).

Wakeley (2008) pointed to reports that the male/female mutation rate ratio may indeed have changed during primate evolution (Goetting-Minesky and Makova 2006) and suggested that the estimate of α used by Patterson et al. ($\alpha = 1.9$, calculated from $\mu_{(X)}/\mu_{(A)} = 0.899$ for H–M) might be too low. Wakeley (2008) suggested that if α was larger in the HC ancestor ($\alpha = 3.7$), the $d_{(X)}/d_{(A)}$ ratio for HC would fit the theoretical expectation from the calculation of Patterson et al.'s. We return to this point below.

Although considering a change in $\mu_{(X)}/\mu_{(A)}$ to be unlikely, Patterson et al. (2006) did not discuss the relative importance of the $T_{(X)}/T_{(A)}$ and $N_{(X)}/N_{(A)}$ ratios as possible causes for the reduced $d_{(X)}/d_{(A)}$ ratio in the HC ancestor (see eq. 8). A difference in speciation time, $T_{(X)} < T_{(A)}$, if established, would provide strong support for the hypothesis of complex speciation with introgression. Introgression is often followed by nonrandom removal or retainment of genomic regions in the recipient species depending on whether the loci are involved in genomic incompatibility (e.g., Noor et al. 2000; Bull et al. 2006; Geraldes et al. 2006). Loci contributing to reproductive isolation tend to be overrepresented on the X chromosome due to the so-called "large X effect" (Dobzhansky 1936). Patterson et al. (2006; 2008) postulated that a complex speciation process involving hybridization after the separation of the 2 species might be the cause for the reduced H–C sequence divergence. An even more complex scenario has been suggested by Mallet (2007, personal communication), in which hybridization occurred in a small local subpopulation, and the chimpanzee X carrying advantageous alleles spread as a wave of advance across the rest of the large human population, whereas almost all the chimpanzee autosomal genome became diluted and lost through meiotic reassortment (except for a very few loci under selection to ameliorate hybrid incompatibilities). Depending on the strength of selection and whether there was enough time for recombination to act, the whole or part of the X chromosome could be affected. The model predicts a chimpanzee origin for the X chromosome and human origin for the autosomes. Nevertheless, this complex speciation model as well as the versions proposed by Patterson et al. (2006, 2008) appears to predict a small τ_{HC} for X instead of a small θ_{HC} for X.

By simultaneously analyzing data from multiple species under a rigorous probabilistic framework, we were able to resolve the factors that contributed to the small $d_{(X)}/d_{(A)}$ for H–C based solely on the data. Our estimates showed that $\tau_{(X)}/\tau_{(A)}$ for H–C was similar to other species pairs, whereas $\theta_{(X)}/\theta_{(A)}$ was unusually small (table 9). Because $\tau_{(X)}/\tau_{(A)}$ is the product of $T_{(X)}/T_{(A)}$ and $\mu_{(X)}/\mu_{(A)}$, we argued that to a good approximation, $T_{(X)} = T_{(A)}$ and $\mu_{(X)}/\mu_{(A)}$ is constant. We drew the following conclusions about the process of human–chimpanzee speciation. 1) The low H–C divergence on X is not explained by current models of hybridization and introgression because these models imply small $\tau_{(X)}/\tau_{(A)}$, not the observed small $\theta_{(X)}/\theta_{(A)}$. 2) Although the data cannot exclude some increase in α in the HC ancestor as postulated by Wakeley (2008), this was not the principal cause of the low divergence on X because we saw no significant drop in $\mu_{(X)}/\mu_{(A)}$ for H–C. 3) The principal reason for the unusually low divergence observed on X between humans and chimpanzees was a major transient reduction in X-linked effective population size (small $N_{(X)}/N_{(A)}$) in the HC ancestor.

The reasons for the small $N_{(X)}$ in the HC ancestor are unknown. In theory, a highly unbalanced sex ratio or high variance in the reproductive success in females could cause $N_{(X)}/N_{(A)} < 0.75$, but to produce $N_{(X)}/N_{(A)} = 0.51$, the female effective population size would need to be close to zero. This explanation thus seems biologically unreasonable. Selection at linked loci may reduce the effective population size at neutral sites under 2 well-established models. In the "hitchhiking" model (Maynard Smith and Haigh 1974), the selective sweep to fixation of a beneficial mutant wipes out standing polymorphism at linked neutral loci. In the "background selection" model (Charlesworth et al. 1993), the steady removal of deleterious mutants has a similar effect in reducing polymorphism at linked neutral loci. However, increased exposure of deleterious recessives on X should lead to a lower frequency of deleterious alleles, and a weakened effect of background selection on X, leading to higher rather than lower $N_{(X)}/N_{(A)}$ for neutral loci (Charlesworth 1996). In a modern population of *Drosophila simulans*, Begun and Whitley (2000) found greatly reduced X chromosome diversity and proposed that positive selection (hitchhiking) was the most likely cause. Theoretical studies by Betancourt et al. (2004) suggest that selective sweeps are indeed more effective in reducing diversity at linked neutral loci on X than on the autosomes. This effect is due to 2 factors. First, advantageous mutations, if they are partially recessive, will have a faster fixation (substitution) rate on X because partially recessive alleles are somewhat shielded from selection on the autosomes but exposed on X (Charlesworth et al. 1987). Second, the shorter sojourn time of a beneficial mutation on X en route to fixation means fewer generations for recombination to occur during a selective sweep (Aquadro et al. 1994). In *Drosophila*, the absence of recombination in males offsets the second effect by increasing the relative recombination rate on X. In mammals, where chromosomes do recombine in males, the relative loss of diversity on X is predicted to be greater (Betancourt et al. 2004), but it is unclear whether selective sweeps can cause such a large reduction in relative population size ($N_{(X)}/N_{(A)}$) as we observe in the human–chimpanzee ancestor.

Supplementary Material

Supplementary tables S1–S4 are available at *Molecular Biology and Evolution* online (http://www.mbe.oxfordjournals.org/).

Acknowledgments

We thank Hiroshi Akashi, Ingo Ebersberger, Arndt von Haeseler, Asger Hobolth, Jim Mallet, Yoko Satta, Mikkel Schierup, Monty Slatkin, and Naoyuki Takahata for discussions and John Wakeley, David Reich, and Nick Patterson for many constructive comments on earlier versions of the manuscript. Part of this article was written while the authors were visiting the Laboratory of Biometrics, University of Tokyo. We thank our host Professor Hirohisa Kishino for his hospitality. This study is supported by a grant from Natural Environment Research Council (UK) to Z.Y. and a Starr Foundation Centennial Scholarship to R.B.

Literature Cited

Aquadro CF, Begun DJ, Kindahl EC. 1994. Selection, recombination, and DNA polymorphism in *Drosophila*. In: Golding B, editor. Non-neutral evolution: theories and molecular data. New York: Chapman & Hall. p. 46–55.

Bailey JA, Carrel L, Chakravarti A, Eichler EE. 2000. Molecular evidence for a relationship between LINE-1 elements and X chromosome inactivation: the Lyon repeat hypothesis. Proc Natl Acad Sci USA. 97:6634–6639.

Bannert N, Kurth R. 2006. The evolutionary dynamics of human endogenous retroviral families. Annu Rev Genomics Hum Genet. 7:149–173.

Barton NH. 2006. Evolutionary biology: how did the human species form? Curr Biol. 16:R647–R650.

Batzer MA, Deininger PL. 2002. Alu repeats and human genomic diversity. Nat Rev Genet. 3:370–379.

Begun DJ, Whitley P. 2000. Reduced X-linked nucleotide polymorphism in Drosophila simulans. Proc Natl Acad Sci USA. 97:5960–5965.

Betancourt AJ, Kim Y, Orr HA. 2004. A pseudohitchhiking model of X vs. autosomal diversity. Genetics. 168: 2261–2269.

Brunet M, Guy F, Pilbeam D, et al. (35 co-authors). 2002. A new hominid from the upper Miocene of Chad, central Africa. Nature. 418:145–151.

Bull V, Beltran M, Jiggins C, McMillan WO, Bermingham E, Mallet J. 2006. Polyphyly and gene flow between non-sibling Heliconius species. BMC Biol. 4:11.

Charlesworth B. 1996. Background selection and patterns of genetic diversity in Drosophila melanogaster. Genet Res. 68:131–149.

Charlesworth B, Coyne JA, Barton NH. 1987. The relative rates of evolution of sex chromosomes and autosomes. Am Nat. 130:113–146.

Charlesworth B, Morgan MT, Charlesworth D. 1993. The effect of deleterious mutations on neutral molecular variation. Genetics. 134:1289–1303.

Chen F-C, Li W-H. 2001. Genomic divergences between humans and other hominoids and the effective population size of the common ancestor of humans and chimpanzees. Am J Hum Genet. 68:444–456.

Consortium TCSaA. 2005. Initial sequence of the chimpanzee genome and comparison with the human genome. 437:69–87.

Dawid AP. 1999. The trouble with Bayes factors. Research Report 202. London: University College London, Department of Statistical Science.

Dobzhansky T. 1936. Studies on hybrid sterility factors in Drosophila pseudoobscura hybrids. Genetics. 21:113–135.

Ebersberger I, Galgoczy P, Taudien S, Taenzer S, Platzer M, von Haeseler A. 2007. Mapping human genetic ancestry (10.1093/molbev/msm156). Mol Biol Evol. 24:2266–2276.

Edgar RC. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. 32:1792–1797.

Ellegren H. 2007. Characteristics, causes and evolutionary consequences of male-biased mutation. Proc R Soc Lond B Biol Sci. 274:1–10.

Fischer A, Pollack J, Thalmann O, Nickel B, Paabo S. 2006. Demographic history and genetic differentiation in apes. Curr Biol. 16:1133–1138.

Gage TB. 1998. The comparative demography of primates: with some comments on the evolution of life histories. Annu Rev Anthropol. 27:197–221.

Geraldes A, Ferrand N, Nachman MW. 2006. Contrasting patterns of introgression at X-linked loci across the hybrid zone between subspecies of the European rabbit (*Oryctolagus cuniculus*). Genetics. 173:919–933.

Goetting-Minesky MP, Makova KD. 2006. Mammalian male mutation bias: impacts of generation time and regional variation in substitution rates. J Mol Evol. 63:537–544.

Goodman M. 1961. The role of immunochemical differences in the phyletic development of human behavior. Hum Biol. 33:131–162.

Haldane JBS. 1935. The rate of spontaneous mutation of a human gene. J Genet. 31:317–326.

Hasegawa M, Kishino H, Yano T. 1985. Dating the human-ape splitting by a molecular clock of mitochondrial DNA. J Mol Evol. 22:160–174.

Hobolth A, Christensen OF, Mailund T, Schierup MH. 2007. Genomic relationships and speciation times of human, chimpanzee, and gorilla inferred from a coalescent hidden Markov model. PLoS Genet. 3:e7.

Hsu F, Kent WJ, Clawson H, Kuhn RM, Diekhans M, Haussler D. 2006. The UCSC known genes. Bioinformatics. 22:1036–1046.

Hudson RR. 1983. Testing the constant-rate neutral allele model with protein sequence data. Evolution. 37:203–217.

Innan H, Watanabe H. 2006. The effect of gene flow on the coalescent time in the human-chimpanzee ancestral population. Mol Biol Evol. 23:1040–1047.

Johnson PLF, Slatkin M. 2008. Accounting for bias from sequencing error in population genetic estimates. Mol Biol Evol. 25:199–206.

Jukes TH. 1987. Transitions, transversions, and the molecular evolutionary clock. J Mol Evol. 26:87–98.

Jukes TH, Cantor CR. 1969. Evolution of protein molecules. In: Munro HN, editor. Mammalian protein metabolism. New York: Academic Press. p. 21–123.

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

Khan H, Smit A, Boissinot S. 2006. Molecular evolution and tempo of amplification of human LINE-1 retrotransposons since the origin of primates. Genome Res. 16:78–87.

Kim SH, Elango N, Warden C, Vigoda E, Yi SV. 2006. Heterogeneous genomic molecular clocks in primates. PLoS Genet. 2:e163.

Kimura M. 1980. A simple method for estimating evolutionary rate of base substitution through comparative studies of nucleotide sequences. J Mol Evol. 16:111–120.

Kimura M, Crow JF. 1964. The number of alleles that can be maintained in a finite population. Genetics. 49:725–738.

Kong X, Murphy K, Raj T, He C, White PS, Matise TC. 2004. A combined linkage-physical map of the human genome. Am J Hum Genet. 75:1143–1148.

Lander ES, Linton LM, Birren B, et al. (251 co-authors). 2001. Initial sequencing and analysis of the human genome. Nature. 409:860–921.

Leakey MG, Ungar PS, Walker A. 1995. A new genus of large primate from the late Oligocene of Lothidok, Turkana District, Kenya. J Hum Evol. 28:519–531.

Li WH, Ellsworth DL, Krushkal J, Chang BH, Hewett-Emmett D. 1996. Rates of nucleotide substitution in primates and rodents and the generation-time effect hypothesis. Mol Phylogenet Evol. 5:182–187.

Lyon MF. 1998. X-chromosome inactivation: a repeat hypothesis. Cytogenet Cell Genet. 80:133–137.

Mallet J. 2007. Hybrid speciation. Nature. 446:279–283.

Maynard Smith J, Haigh J. 1974. The hitch-hiking effect of a favorable gene. Genet Res (Camb). 23:23–35.

Mikkelsen TS, Hillier LW, Eichler EE, Zody MC, Jaffe DB, Yang S-P, Enard W, Hellmann I, Lindblad-Toh K, Altheide TK. (64 co-authors). 2005. Initial sequence of the chimpanzee genome and comparison with the human genome. 437:69–87.

Miyata T, Hayashida H, Kuma K, Mitsuyasu K, Yasunaga T. 1987. Male-driven molecular evolution: a model and nucleotide sequence analysis. Cold Spring Harb Symp Quant Biol. 52:863–867.

Nei M, Graur D. 1984. Extent of protein polymorphism and the neutral mutation theory. Evol Biol. 17:73–118.

Noor MA, Johnson NA, Hey J. 2000. Gene flow between Drosophila pseudoobscura and D. persimilis. Evolution. 54:2174–2175; discussion 2176–2177.

Patterson N, Richter DJ, Gnerre S, Lander ES, Reich D. 2006. Genetic evidence for complex speciation of humans and chimpanzees. Nature. 441:1103–1108.

Patterson N, Richter DJ, Gnerre S, Lander ES, Reich D. 2008. Complex speciation of humans and chimpanzees (Reply to Wakeley). Nature. 452:E4.

Ptak SE, Hinds DA, Koehler K, Nickel B, Patil N, Ballinger DG, Przeworski M, Frazer KA, Paabo S. 2005. Fine-scale recombination patterns differ between chimpanzees and humans. Nat Genet. 37:429–434.

Quentin Y. 1992. Origin of the Alu family: a family of Alu-like monomers gave birth to the left and the right arms of the Alu elements. Nucleic Acids Res. 20:3397–3401.

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.

Sachidanandam R, Weissman D, Schmidt SC, et al. (38 co-authors). 2001. A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. Nature. 409:928–933.

Satta Y, Takahata T, Schnbach C, Gutknecht J, Klein J. 1991. Calibrating evolutionary rates at major histocompatibility complex loci. In: Klein J, Klein D, editors. Molecular evolution of the major histocompatibility complex. Heidelberg (Germany): Springer. p. 51–62.

Satta Y, Hickerson M, Watanabe H, O'hUigin C, Klein J. 2004. Ancestral population sizes and species divergence times in the primate lineage on the basis of intron and BAC end sequences. J Mol Evol. 59:478–487.

Senut B, Pickford M, Gommery D, Mein P, Cheboi K, Coppens Y. 2001. First hominid from the Miocene (Lukeino Formation, Kenya). C R Acad Sci II. 332:137–144.

Smit AF. 1993. Identification of a new, abundant superfamily of mammalian LTR-transposons. Nucleic Acids Res. 21: 1863–1872.

Steiper ME, Young NM. 2006. Primate molecular divergence dates. Mol Phylogenet Evol. 41:384–394.

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

Takahata N. 1986. An attempt to estimate the effective size of the ancestral species common to two extant species from which homologous genes are sequenced. Genet Res. 48:187–190.

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

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.

Taylor J, Tyekucheva S, Zody M, Chiaromonte F, Makova KD. 2006. Strong and weak male mutation bias at different sites in the primate genomes: insights from the human-chimpanzee comparison. Mol Biol Evol. 23:565–573.

Tenesa A, Navarro P, Hayes BJ, Duffy DL, Clarke GM, Goddard ME, Visscher PM. 2007. Recent human effective population size estimated from linkage disequilibrium. Genome Res. 17:520–526.

Wakeley J. 2008. Complex speciation of humans and chimpanzees. Nature. 452:E3–E4.

Yang Z. 1994. Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: approximate methods. J Mol Evol. 39:306–314.

Yang Z. 1997a. PAML: a program package for phylogenetic analysis by maximum likelihood. Comput Appl Biosci. 13: 555–556.

Yang Z. 1997b. On the estimation of ancestral population sizes. Genet Res. 69:111–116.

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

Yi S, Ellsworth DL, Li WH. 2002. Slow molecular clocks in Old World monkeys, apes, and humans. Mol Biol Evol. 19: 2191–2198.

Yu N, Jensen-Seaman MI, Chemnick L, Ryder O, Li WH. 2004. Nucleotide diversity in gorillas. Genetics. 166:1375–1383.

Yu N, Zhao Z, Fu YX, Sambuughin N, Ramsay M, Jenkins T, Leskinen E, Patthy L, Jorde LB, Kuromori T, Li W.-H. 2001. Global patterns of human DNA sequence variation in a 10-kb region on chromosome 1. Mol Biol Evol. 18:214–222.

Rasmus Nielsen, Associate Editor

Accepted June 29, 2008