Estimating Ancestral Population Sizes and Divergence Times

Jeffrey D. Wall[1]

Department of Human Genetics, University of Chicago, Chicago, Illinois 60637

Manuscript received February 12, 2002
Accepted for publication October 14, 2002

ABSTRACT

This article presents a new method for jointly estimating species divergence times and ancestral population sizes. The method improves on previous ones by explicitly incorporating intragenic recombination, by utilizing orthologous sequence data from closely related species, and by using a maximum-likelihood framework. The latter allows for efficient use of the available information and provides a way of assessing how much confidence we should place in the estimates. I apply the method to recently collected intergenic sequence data from humans and the great apes. The results suggest that the human-chimpanzee ancestral population size was four to seven times larger than the current human effective population size and that the current human effective population size is slightly >10,000. These estimates are similar to previous ones, and they appear relatively insensitive to assumptions about the recombination rates or mutation rates across loci.

THE effective population size (N_e) of a species has a direct effect on the amount and the pattern of DNA sequence variation. Researchers have therefore used sequence polymorphism data to estimate N_e (*e.g.*, Kreitman 1983; Takahata 1993; Nachman and Crowell 2000). The amount of observed diversity can be used to estimate the population mutation parameter $\theta = 4N_e\mu$, and the per generation mutation rate μ can be estimated either directly (Harada *et al.* 1993; Giannelli *et al.* 1999) or indirectly (*e.g.*, Kimura 1983; Satta *et al.* 1993; Kumar and Hedges 1998; Nachman and Crowell 2000) from divergence data (given assumptions about the divergence date and the average generation time).

Most estimates of N_e for humans are ~10,000–15,000 (*e.g.*, Takahata 1993; Harding *et al.* 1997). While there are many possible reasons why the effective population size may be quite different from the census population size (Caballero 1994), it remains surprising that the human N_e is so low, especially given humans' large range over the past 1–2 million years (MY; *e.g.*, Swisher *et al.* 1994; Gabunia and Vekua 1995). In particular, great ape species historically have had much smaller ranges, but have N_e two to three times larger than the human N_e. Did some event associated with the founding of the genus Homo (Takahata 1993) or some other particular event in human history lead to a sharp reduction in effective population size? It is difficult to answer this without knowing how N_e has varied over evolutionary time. Recently, progress has been made in estimating the effective population size of the population directly ancestral to two extant daughter species.

A few main methods exist for estimating N_a (see Takahata and Satta 2002 for a more in-depth discussion). (We refer to the ancestor's population size as the "ancestral N_e," or N_a.) One method, referred to as the trichotomy method, requires orthologous sequence data from three closely related species (Nei 1987; Wu 1991). This approach uses a single orthologous sequence from each of three species and assumes a simple model of population history where at fixed times different species become isolated with no further admixture (*cf.* Hey 1994). Random mating is assumed within each population. If the time between the two speciation events is small, the gene tree for a particular region will not always match the species tree (Nei 1987; see Figure 1). The probability that this happens depends in part on N_a for the ancestral population of species 1 and 2. In particular, a necessary (but not sufficient) condition for the gene tree and the species tree to be incompatible is that the species 1 and 2 lineages do not coalesce between the two speciation events. If N_a is larger, the probability of a common ancestor before time T_2 is reduced, leading to a greater chance that the gene tree and the species tree do not match. The trichotomy method uses orthologous data from many unlinked loci, infers the gene tree at each locus, calculates the proportion of loci where the inferred gene tree does not match the species tree, and then uses this proportion to estimate N_a. Application of the trichotomy method to human and great ape sequence data has led to estimates of N_a (for the human-chimpanzee ancestral population) substantially larger than the current N_e for humans (Ruvolo 1997; Chen and Li 2001; Takahata and Satta 2002). Chen and Li (2001) estimate, for example, N_a = 52,000–96,000,

[1]*Address for correspondence:* Department of Human Genetics, 920 E. 58th St.—CLSC 507, Chicago, IL 60637.
E-mail: jwall@genetics.bsd.uchicago.edu

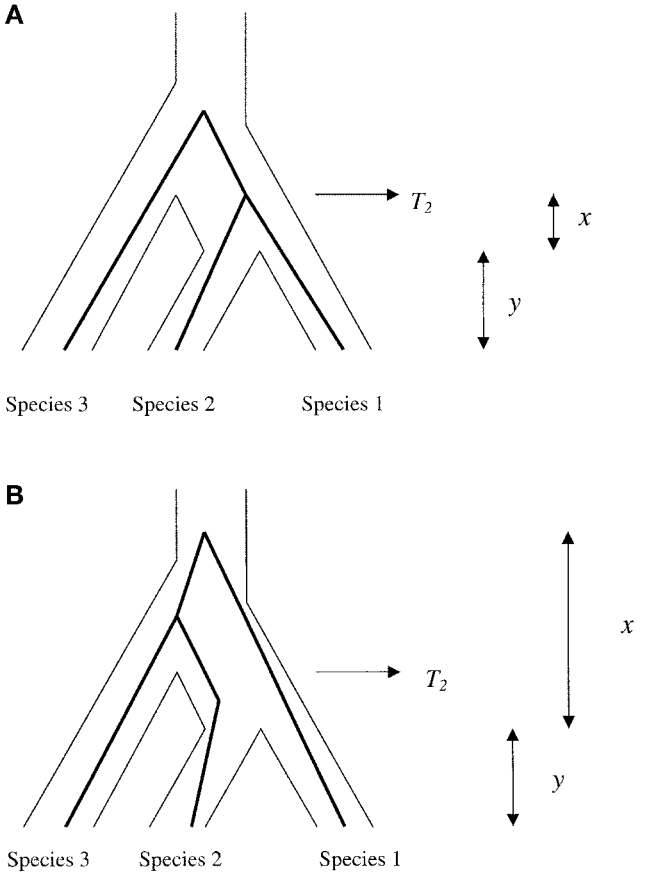

FIGURE 1.—Two possible gene trees given a particular species tree. T_2 is the time when the first speciation event occurs. In A, the gene tree and species tree are compatible, while in B they are incompatible. Divergence between single orthologous sequences from two species (species 1 and 2) consists of two parts: time when the species are separated (y) and the time when the two sequences segregate in the ancestral population (x).

or roughly five to nine times larger than the current human N_e.

Another method for estimating N_a requires divergence data from two or three species (TAKAHATA 1986; TAKAHATA *et al.* 1995; YANG 1997, 2002). Here, I describe the two-species method, because that is what is generally used. Given two orthologous sequences, one each from a pair of species, they will coalesce at some time that predates the species divergence time (see Figure 1). For the autosomes, the time spent in the ancestral population before coalescence (x in Figure 1) is exponentially distributed, with mean $2N_a g$ (where g is the average generation time and N_a is the diploid ancestral N_e). In contrast, the postspeciation branch lengths (y in Figure 1) are fixed. Given data from multiple unlinked loci and assumptions about g and μ, one can use maximum likelihood to jointly estimate the speciation time and N_a (TAKAHATA *et al.* 1995; YANG 1997). The general idea is that large values of N_a correspond to greater variability in the coalescence time of two orthologous sequences and thus greater variance in the observed divergences across loci. Using human and chimpanzee divergence data, estimates of the human-chimpanzee N_a are ~5–10 times the current N_e (TAKAHATA and SATTA 1997; TAKAHATA 2001).

Finally, two other methods require intraspecific polymorphism data from two species and use either a moment-based (WAKELEY and HEY 1997) or a maximum-likelihood (NIELSEN and WAKELEY 2001) approach to estimate model parameters (including in particular N_a and the divergence time). Both of these methods are well suited for species that have diverged relatively recently, but less so for species such as humans and chimpanzees that share very little ancestral polymorphism. In any case, at the present they cannot be used to estimate the human-chimp N_a because of a lack of chimpanzee polymorphism data.

Large estimates of the human-chimpanzee N_a are concordant with a study of *Mhc* that used the high levels of diversity there to estimate a long-term (*i.e.*, over the past 10–20 million years) average effective population size of $\sim 10^5$ (TAKAHATA 1991). However, it should be noted that the large estimates of N_a are difficult to reconcile with human-chimpanzee divergence times estimated from molecular data. Most recent estimates of the divergence time fall between 4 and 6 million years ago (MYA; *e.g.*, HORAI *et al.* 1995; EASTEAL and HERBERT 1997; KUMAR and HEDGES 1998; KUMAR and SUBRAMANIAN 2002). These estimates are for a single human and a single chimpanzee sequence; they reflect both divergence between species and segregation in the ancestral population (x and y in Figure 1). If N_a is large, then x must be large; if x is large and $x + y$ is fixed, then y must be small. Suppose, for example, that $(x + y) = 5.5$ MY, as estimated by KUMAR and HEDGES (1998), and that $N_a = 52{,}000$–$96{,}000$ (*cf.* CHEN and LI 2001). Then, if the average generation time is 20 years, y would be 1.7–3.4 MY. If the average generation time were 25 years (see DISCUSSION), then $y = 0.7$–2.9 MY. These estimates postdate many well-documented australopithecine fossils and are therefore dubious estimates of the time since speciation.

One possible explanation is that N_a has been consistently overestimated. Indeed, both the trichotomy method and the two-species maximum-likelihood method have been criticized (HUDSON 1992; TAKAHATA *et al.* 1995; SATTA *et al.* 2000; TAKAHATA and SATTA 2002), and it is not clear how accurate the estimates are. For one, the trichotomy method assumes one already knows the time between the two speciation events, but this is generally not known *a priori*. Furthermore, it assumes one can correctly infer the phylogeny for any particular locus. Errors in phylogenetic inference arise when analyzing actual data, and the whole endeavor does not make sense in the presence of intragenic recombination (*cf.* NORDBORG 2001). With three closely related species,

the true phylogenies for nearby sites are not always the same. So, when loci are analyzed, they are often an amalgamation of sites with different phylogenies. Trying to infer a single phylogeny from such data is clearly not appropriate (Satta *et al.* 2000). Even if the problems of recombination and phylogenetic reconstruction were ignored, the trichotomy method does not make an efficient use of the available information. Data from each locus are summarized into a single binary variable, depending on whether the inferred locus phylogeny agrees or disagrees with the species tree.

The maximum-likelihood methods are more rigorous and efficient, but they too have two main drawbacks. As with the trichotomy method and the method of Nielsen and Wakeley (2001), intragenic recombination is ignored. In addition, the methods of Takahata *et al.* (1995) are highly sensitive to variation in μ among loci. The problem is that variation in μ leads to greater variance in observed divergences across loci, which inflates the estimate of N_a. Although variation in μ can be explicitly modeled in the analyses (Yang 1997; Takahata and Satta 2002), it is difficult to know whether a particular model of rate variation is appropriate, especially given data from only two species (but see Yang 2002).

In this article, I present a new method for estimating N_a. The method requires orthologous sequence data from three or more species (two plus one or more outgroups) and jointly estimates N_a and species divergence times using a summary maximum-likelihood approach. Unlike the previous maximum-likelihood methods, intragenic recombination is incorporated, and likelihoods are estimated from coalescent simulations. Also, the model can account for variation in mutation rates across loci. Although the method can be used on data from any taxonomic group (as long as at least one outgroup species is available), I concentrate here on analyzing human and great ape sequence data. The maximum-likelihood framework allows for the estimation of confidence intervals; this, along with a more realistic model, allows us to assess with greater rigor whether the human-chimp N_a was much larger than the current human N_e, as previous studies have claimed. I apply the method to the orthologous data from 53 intergenic regions reported in Chen and Li (2001) and generate both point estimates and approximate confidence intervals for N_a and species divergence times. Intergenic sequence data are preferable to data from genes (even synonymous sites or introns) because they are less likely to have been affected by natural selection at closely linked sites.

METHODS

I describe the model in which there are orthologous sequence data from four species. The case in which there are three species (or five or more) follows analogously.

Suppose we have four species with a known phylogeny. We assume a null model of speciation (*cf.* Hey 1994; Satta *et al.* 2000) whereby a panmictic ancestral population splits at a fixed time into two panmictic descendant populations, with no subsequent migration between the descendant populations. The scaled mutation and recombination parameters are θ ($= 4N_h\mu$) and ρ ($= 4N_h r$), where μ is the mutation rate per site per generation and r is the recombination rate per site per generation. Label the species H, C, G, and O, with current diploid effective population sizes N_h, N_c, N_g, and N_o, respectively. Suppose H and C split at time T_1, H and G at time T_2, and H and O at time T_3, with $T_1 < T_2 < T_3$ (see Figure 2). T_1, T_2, and T_3 are scaled in units of $4N_h$ generations. From time T_1 until T_3 both the H-C ancestral population and the H-C-G ancestral population have effective size N_a, while the H-C-G-O ancestral population has effective size N_o. The results are similar if the latter ancestral population has effective size N_a (results not shown). Finally, define n_s as the number of contiguous nucleotide sites in the simulation. There are a total of 11 parameters in the model, listed in Table 1. We assume that the generation time and mutation rate do not vary across species. These assumptions are reasonable when the species considered have similar life-history traits and are closely related. There are three possible (unrooted) gene trees, with H and C, H and G, or C and G as sibling species.

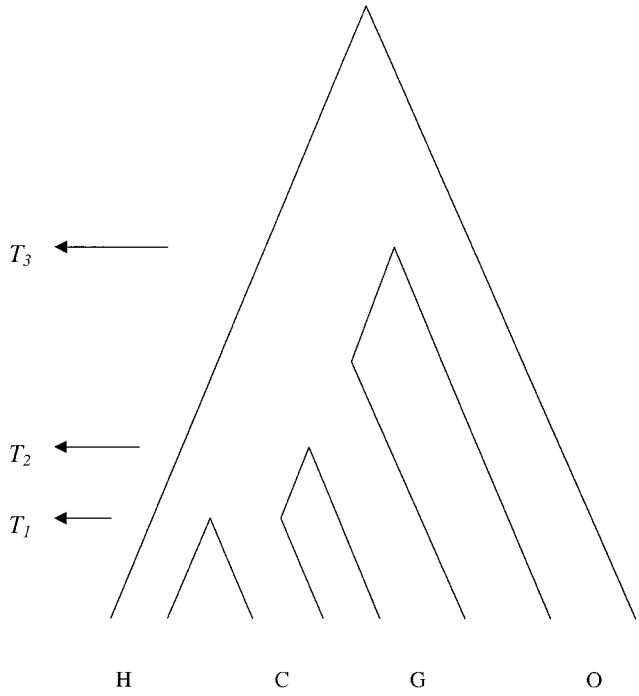

Figure 2.—Model of species history considered. There are four species with known branching order, labeled H, C, G, and O. The three speciation events (starting from the present) occur at times T_1, T_2, and T_3, where time is scaled in units of $4N_h$ generations. See methods for more details.

Now, suppose we have a single orthologous sequence

TABLE 1

Model parameters

Parameter	Definition
θ	$= 4N_h\mu$, where μ is the mutation rate per site per generation.
ρ	$= 4N_h r$, where r is the recombination rate per site per generation.
N_h	The current effective population size of humans.
N_c	The current effective population size of chimpanzees.
N_g	The current effective population size of gorillas.
N_o	The current effective population size of orangutans.
N_a	The ancestral population size for the human-chimp and human-gorilla ancestors.
T_1	The time (in units of $4N_h$ generations) of the human-chimp split.
T_2	The time (in units of $4N_h$ generations) of the human-gorilla split.
T_3	The time (in units of $4N_h$ generations) of the human-orang split.
n_s	Number of contiguous sites in a locus.

from each species. For each site that is "segregating" (*i.e.*, is not identical across all species), we can infer that one or more mutations happened on certain branches in the unrooted tree. We do this assuming the fewest number of mutations that can explain the data. For example, if the H, C, G, and O sequences have A, G, A, and A, respectively, then we infer that a mutation happened on the branch leading to species C. All biallelic segregating sites fall into seven categories, resulting from mutations on seven different branches of an unrooted tree. These seven branches have H, C, G, O, HC, HG, or CG as descendants and are referred to as the seven types of branches. Note that for any particular gene tree there are only five possible branches, four external ones (with a single species as a descendant) and one internal one. Any site may have one of three possible gene trees, leading to seven possible branches over all possible gene trees (the four external branches that are common to each gene tree and one internal branch from each gene tree). For sites with three segregating nucleotides, we assume that the two species with the same base share the ancestral state and that the two other bases each arose from a single mutation. The Chen and Li (2001) data do not contain any sites where each species has a different nucleotide, so we do not consider this possibility.

The sequence data for the 53 intergenic regions reported in Chen and Li (2001) were kindly provided by the authors and aligned by eye. All indels were excluded. From the remaining sequence, we count the inferred number of mutations that happened on each of the seven branch types. (No distinction is made between transitions and transversions.) For a given region, denote these numbers of inferred mutations as $\boldsymbol{b} = (b_1, b_2, \ldots b_7)$. For given values of $\boldsymbol{M} = (\theta, \rho, N_h, N_c, N_g, N_o, N_a, T_1, T_2, T_3, n_s)$ we estimate the likelihood of observing the vector $\boldsymbol{b}$ using Monte Carlo simulations.

The population model in Figure 2 is simulated using a modification of the coalescent with recombination (Hudson 1983). For each site in each replicate, we classify all the branches in the genealogy as one of the seven types of branches (*i.e.*, with descendants H, C, G, O, HC, HG, or CG in the unrooted tree). Since mutations happen at rate μ per site per generation, we can tabulate from the total branch lengths the expected number of mutations that lie on each of the types of branches. For the jth replicate, denote these expected values as $\boldsymbol{B}_j = (B_1, B_2, \ldots B_7)$. The probability of observing $\boldsymbol{b}$ given $\boldsymbol{B}_j$ is then

$$\Pr(\mathbf{b}|\mathbf{B}_j) = \prod_{i=1}^{7} e^{-B_i} \frac{B_i^{b_i}}{b_i!}.$$

To estimate the likelihood of $\boldsymbol{M}$, we just average this probability over many replicates,

$$\mathrm{lik}(\boldsymbol{M}|\boldsymbol{b}) \propto \Pr(\boldsymbol{b}|\boldsymbol{M}) \approx \frac{1}{x} \sum_{j=1}^{x} \Pr(\boldsymbol{b}|\boldsymbol{B}_j),$$

where x is large. The Chen and Li (2001) data consist of two sequences from each species (a single diploid sequence), while the method requires a single sequence. Intraspecies polymorphism may add to the b_i values, depending on which chromosome is considered. In these cases, we take the average likelihood over the different possible b_i values.

The above equation describes how to estimate the likelihood of $\boldsymbol{M}$ for a single locus. Define $\boldsymbol{M}'$ as a vector containing the first 10 values of $\boldsymbol{M}$. Estimation of the likelihood of $\boldsymbol{M}'$ over multiple loci is straightforward. Given a collection of k loci, define $\{\boldsymbol{M}_i\}_{i=1}^{k}$ as a collection of corresponding $\boldsymbol{M}$ vectors, where the $\boldsymbol{M}_i$ are identical except for n_s (which is calculated for each locus). Define $\boldsymbol{b}_i$ as the vector $\boldsymbol{b}$ for the ith locus. Then, since unlinked loci are evolutionarily independent, we can estimate the likelihood of $\boldsymbol{M}'$ over multiple loci simply by taking the product of the individual $\mathrm{lik}(\boldsymbol{M}|\boldsymbol{b})$ estimates:

$$\mathrm{lik}(\boldsymbol{M}') = \prod_{i=1}^{k} \mathrm{lik}(\boldsymbol{M}_i|\boldsymbol{b}_i).$$

We have taken the approach of summarizing the data

by $\boldsymbol{b}$ before performing maximum likelihood. Summary-likelihood methods have been quite useful in other situations (*e.g.*, Weiss and Von Haeseler 1998; Wall 2000; Fearnhead and Donnelly 2002) and are generally computationally much simpler than full-likelihood methods. For the case of estimating N_a, a full-likelihood approach including intragenic recombination does not look to be computationally feasible at this time.

Of the 11 parameters that make up the model $\boldsymbol{M}$, only 9 can freely vary. n_s is fixed from the actual data, while N_h is relevant only indirectly; it turns out that the simulations use only the ratios of the effective population sizes (*i.e.*, N_a/N_h, N_c/N_h, etc.), not their actual values. The actual N_h comes into play when interpreting the simulation results (*e.g.*, translating from scaled time to actual time). Ideally, one would like to let θ, ρ, N_c/N_h, N_g/N_h, N_o/N_h, N_a/N_h, T_1, T_2, and T_3 vary freely and determine which combination of parameter values maximizes the likelihood of observing the actual data. However, this is computationally prohibitive, so we fix those values for which we have prior information and let the others vary: N_a/N_h, T_1, T_2, and T_3 vary freely (at increments of 1.0, 0.25, 0.5, and 1.0, respectively), and we consider the following four schemes for the other parameters:

Model 1: $\theta = \rho = 0.001$/bp; $N_c = N_g = N_o = 3N_h$.
Model 2: θn_s for each locus is proportional to the total inferred number of mutations (across all four species), and the average θ/bp (over all loci) is 0.001; $\rho = 0.001$/bp; $N_c = N_g = N_o = 3N_h$.
Model 3: Same as model 2, but all CpG sites were excluded, and the average θ/bp (over all loci) is 0.00075.
Model 4: $\theta = 0.001$/bp; $\rho = 0.002$/bp; $N_c = N_g = N_o = 3N_h$.
Model 5: $\theta = \rho = 0.001$/bp; $N_c = N_g = N_o = 6N_h$.

θ can be easily estimated from human sequence polymorphism data (*e.g.*, Watterson 1975). Putatively neutral sites from recent resequencing studies of human variation suggest that $\theta = 0.001$/bp is a good ballpark figure for the autosomes and that roughly one-fourth of all segregating mutations occur at CpG sites (*e.g.*, Nachman and Crowell 2000; Przeworski *et al.* 2000; Templeton *et al.* 2000; Ebersberger *et al.* 2001; Frisse *et al.* 2001). Model 2 tests how sensitive the results are to variation in θ across loci by taking the same average θ as model 1, but assuming that θn_s for each locus is proportional to the observed number of inferred mutations. (This is equivalent to estimating θ using Watterson 1975, assuming an average of $\theta = 0.001$/bp.) The genome-wide average rate of crossing over in humans is $r = 1.3 \times 10^{-8}$/bp (Yu *et al.* 2001). If $N_h \approx 10^4$, then $\rho \approx 5.2 \times 10^{-4}$/bp. We take slightly larger ρ values to account for the unknown contribution of gene conversion to overall rates of recombination (see, *e.g.*, Frisse *et al.* 2001; Przeworski and Wall 2001). Finally, levels of nonhuman great ape diversity seem to be substantially higher than human diversity levels (Deinard and Kidd 1999; Kaessmann *et al.* 1999, 2001), but not enough data have been gathered to accurately estimate N_c/N_h, N_g/N_h, or N_o/N_h. We have chosen values that might plausibly reflect the total species diversity in chimps, gorillas, and orangs. Models 3–5 were chosen to explore the sensitivity of the results to the presence of hypermutable CpG sites, assumptions about the recombination rate, and assumptions about great ape population sizes, respectively.

In addition to using the parameter combination that maximizes the likelihood as a point estimate, it would be useful to determine how much confidence we should place in the estimated values. To get a sense of how the likelihood varies as a function of T_1, for example, I calculate the (approximate) profile likelihood:

$$\mathrm{lik}_{T_1}(T) = \sup \prod_{i=1}^{k} \mathrm{lik}(\boldsymbol{M}_i|\boldsymbol{b}_i,\ T_1 = T).$$

Approximate 95% confidence intervals are found by using the standard χ^2 approximation for the likelihood-ratio statistic $2 \ln(L_0/L_1)$ (where L_0 is the maximum likelihood and L_1 is the profile likelihood at an alternative point). The likelihood functions calculated are not true profile likelihoods, since some of the nuisance parameters are not allowed to vary freely. So, it is not clear whether the standard χ^2 approximation is appropriate. Approximate profile likelihoods are calculated for N_a/N_h and T_1, and linear interpolation is used to estimate the log-likelihood for parameter values that are not directly estimated by simulation.

To verify the accuracy of the method, I run coalescent simulations with known T_1, T_2, T_3, and N_a/N_h values; then, I use the new method on the simulated data to estimate parameters and to compare the estimated values with the actual ones. These simulations modeled 50 loci of 500 bp each, with $\theta = \rho = 0.001$/bp, $N_c = N_g = N_o = 3N_h$, $T_1 = 5.0$, $T_2 = 8.0$, $T_3 = 14.0$, and $N_a/N_h = 5.0$. The parameter values were chosen to roughly match both the Chen and Li (2001) data and our *a priori* knowledge about species divergence times and ancestral population sizes. Five replicates were run; I analyzed each one under the assumptions of model 1 (see above). Note that this assumes an idealized situation, where the nuisance parameters are known exactly.

All programs were written in C and are available from the author on request. A total of 5×10^4 replicates were run for each model and parameter combination. To give a sense of the computational efficiency, the total simulations took 5 months to run on a pair of 1.7 GHz Pentium 4 processors.

RESULTS

The maximum-likelihood estimates for T_1, T_2, T_3, and N_a/N_h are presented in Table 2. The estimates across the five models are broadly similar; all of them estimate an ancestral population size five to six times larger than

TABLE 2

Parameter estimates and confidence intervals for the Chen and Li (2001) data

	MLE[a]			Intervals[b]		
	T_1	T_2	T_3	N_a/N_h	T_1	N_a/N_h
Model 1	3.5	5.0	12.0	6.0	2.8–4.3	4.3–7.0
Model 2	4.0	5.5	13.0	5.0	3.0–4.5	3.8–6.5
Model 3	4.0	5.5	13.0	5.0	2.9–4.8	3.5–6.5
Model 4	3.5	5.0	12.0	6.0	2.7–4.2	4.3–7.1
Model 5	3.5	5.0	12.0	6.0	2.8–4.3	4.3–7.0

[a] MLE, maximum-likelihood estimate. Parameter values that maximize the likelihood of the data are shown.

[b] Approximate 95% confidence intervals. See methods for details.

the current human effective population size, in keeping with previous studies (Takahata and Satta 1997; Chen and Li 2001; Takahata 2001). The estimates of T_1, the human-chimpanzee divergence time, are also roughly in line with expectations. If we assume that $g = 25$ years and $N_h = 10^4$ (or that $g = 20$ years and $N_h = 12{,}500$), then these estimates range from 3.5 to 4.0 MYA. In contrast, the paleontological record suggests that uniquely human ancestors were around at least 4–4.5 MYA (White *et al.* 1994; Leakey *et al.* 1995) and perhaps much earlier (Haile-Selassie 2001; Brunet *et al.* 2002). This disparity can easily be reconciled if both the average generation time and the current human effective population size are on the larger side of previous estimates (*e.g.*, $g = 25$ years and $N_h = 15{,}000$). Given our uncertainty in parameter estimates, these values are quite plausible. If instead we were to assume the generation time and species divergence time were known, then we could use the results to estimate the current human effective population size. If $T_1 = 6$ MYA and $g = 25$ years, then the point estimates of N_h range from 15,000 to 17,100. The other species divergence times are also on the recent side; assuming once again that $g = 25$ years and $N_h = 10{,}000$, the estimated human-gorilla divergence time ranges from 5.0 to 5.5 MYA, while the estimated human-orangutan divergence time ranges from 12 to 13 MYA.

To assess how much confidence we should place in the point estimates, I calculated approximate profile-likelihood curves and estimated ~95% confidence intervals. The intervals for T_1 and N_a/N_h are listed in Table 2. For both T_1 and N_a/N_h the intervals are quite narrow, which suggests that the estimates are precise. All four models exclude $N_a/N_h \leq 3.5$ and $N_a/N_h \geq 7.1$ from the approximate confidence intervals. For T_1, the lower boundaries range from 2.7 to 3.0 and the upper boundaries range from 4.2 to 4.8. If as before we take $g = 25$ years and $N = 10^4$, the upper boundaries range from 4.2 to 4.8 MYA; these times are still more recent than the paleontological record would suggest. As mentioned above, a small increase in N_h is sufficient to reconcile the time estimates with the paleontological record. Figure 3 shows the profile-likelihood functions of N_a/N_h and T_1 for model 2. The curves quickly become quite steep, suggesting that the range of plausible values is not that large. So, even if the approximate confidence intervals were nonconservative, it is likely that conservative ones would not differ much from the intervals listed in Table 2. The corresponding likelihood curves for the other models are qualitatively similar to those in Figure 3.

To verify the accuracy of the method, I applied it on five simulated data sets with known parameter values (see methods). Each one had actual values of $T_1 = 5.0$, $T_2 = 8.0$, $T_3 = 14.0$, and $N_a/N_h = 5$. The estimated parameter values, along with the confidence intervals for T_1 and N_a/N_h, are given in Table 3. The means of the parameter estimates are 5.0, 8.1, 14.0, and 4.8 for T_1, T_2, T_3, and N_a/N_h, respectively, which suggests that the method has no or low bias. In addition, the confidence intervals for T_1 and N_a/N_h contain the true value all five times. Due to the large computational burden, it was not possible to run enough replicates to accurately estimate the coverage properties of the confidence intervals.

Comparing the different rows in Table 2 can give us some idea of how sensitive the results are to assumptions about the nuisance parameters (*i.e.*, θ, ρ, N_c/N_h, N_g/N_h, and N_o/N_h). Since the results from all of the models are very similar, it appears that the particular assumptions made do not appear to be very important. In particular, unlike the two-species maximum-likelihood method of Takahata *et al.* (1995), the results do not seem to be very sensitive to variation in mutation rates across loci. This may be due to the information about locus-specific mutation rates contained in the outgroup species or because the actual data have very little variation in mutation rates across loci.

DISCUSSION

Estimating ancestral population sizes has been an active research area for several years. The work presented here improves on previous efforts by explicitly incorporating intragenic recombination (see also Satta *et al.* 2000) and by efficiently utilizing data from outgroup species. The estimates of the human-chimpanzee N_a are five to six times larger than the current human effective population size (see Table 2). Although most previous studies came to similar conclusions, it was not clear how much confidence to place in these estimates because of unrealistic assumptions, such as no recombination or no variation in mutation rates (Takahata and Satta 2002). The narrow confidence intervals and simulation results presented here (Tables 2 and 3) provide additional evidence that ancestral population sizes were substantially larger than the current human effective population size.

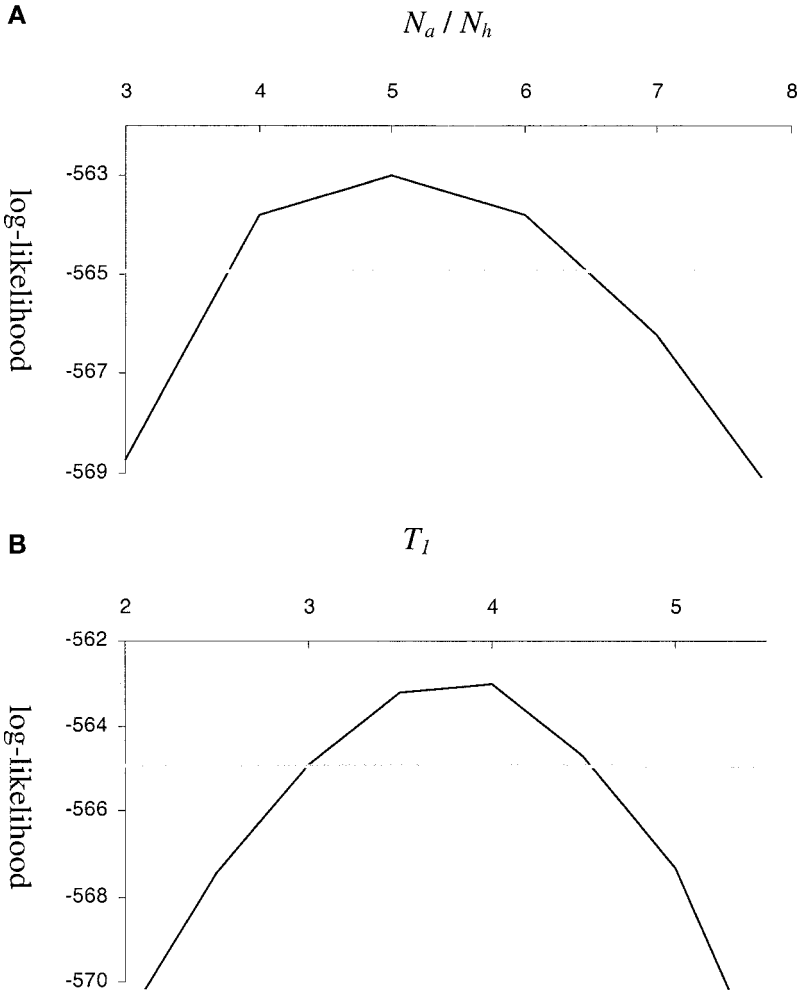

FIGURE 3.—Approximate profile-likelihood curves under model 2. (A) The curve for N_a/N_h; (B) the curve for T_1. In both cases, the *y*-axis is the maximal log-likelihood given a particular value of the parameter (see METHODS for details). The shaded horizontal line shows the cutoff for the ~95% confidence intervals.

One recent study that came to a different conclusion (namely, that N_a is roughly as small as N_h) incorporated variation in mutation rates across loci but not intragenic recombination (YANG 2002). Recombination tends to decrease the variance in estimated branch lengths across loci; because of this, models that assume no recombination tend to underestimate N_a (TAKAHATA and SATTA 2002). Further work must be done to quantify how model assumptions (both here and in other studies) affect estimates of the ancestral population size.

Although this application focuses on the human-chimpanzee N_a, the same method can be used to estimate N_a from other taxa, as long as there are orthologous sequence data from three or more species (including at least one outgroup) at multiple unlinked loci. Below, I discuss issues that might affect the general applicability of the method.

Likelihood model: One possible criticism of the model is that the relative locations of the segregating mutations are ignored. However, this is not likely to be very important, since the number and the pattern of segregating mutations are far more informative. Incorporating the segregating site locations may lead to narrower confidence intervals and more accurate estimation of the likelihood function, but excluding them is not expected to bias the results in either direction. Given the results, it does not seem to be worth the substantial computational burden to consider the full-likelihood model.

Mutational model: The mutational model that was adopted makes no distinction between transitions and transversions and assumes the mutation rate at each site in a locus is the same. However, some sites have higher mutation rates than others (NACHMAN and CROWELL 2000; TEMPLETON *et al.* 2000), which would increase the number of sites experiencing multiple mutations. Those multiply hit sites with fewer than three segregating nucleotides would then be misclassified by the model. In primates, the transition rate away from CpG sites is thought to be elevated by more than an order of magnitude due to methylated-cytosine mutagenesis (*e.g.*, JONES *et al.* 1992; GIANNELLI *et al.* 1999). To test whether homoplasies from multiply hit CpG sites affected the parameter estimates, I reran model 2 excluding all CpG sites (*i.e.*, model 3). Both the maximum-likelihood estimate and the shape of the profile-likelihood curves are almost identical (Table 2; results not shown), suggesting that

TABLE 3

Parameter estimates and confidence intervals for simulated data

	MLE[a]				Intervals[b]	
	T_1	T_2	T_3	N_a/N_h	T_1	N_a/N_h
True value:[c]	5.0	8.0	14.0	5.0		
Trial 1	5.0	8.0	15.0	4.0	4.2–5.7	3.2–5.4
Trial 2	4.5	8.0	14.0	5.0	3.5–5.5	3.5–6.4
Trial 3	4.5	7.5	14.0	5.0	3.4–5.2	3.9–6.7
Trial 4	5.0	8.0	14.0	5.0	4.3–6.1	3.5–6.2
Trial 5	6.0	9.0	13.0	5.0	4.8–6.7	3.8–6.9

All trials were analyzed under model 1. See METHODS for full details.
[a] Parameter values that maximize the likelihood of the data.
[b] Approximate 95% confidence intervals. See METHODS for details.
[c] Parameter values used in the simulations.

the results presented here are relatively insensitive to the effects of multiple mutations at CpG sites. Calculations suggest that other proposed sites with elevated mutation rates, such as mononucleotide runs or DNA polymerase α-arrest sites (KRAWCZAK and COOPER 1991; TEMPLETON *et al.* 2000), are too rare to appreciably increase the expected number of homoplasies (results not shown). For studies of species with larger levels of divergence, the effects of homoplasies may be a more serious concern. Future work will concentrate on implementing a finite-site mutation model in the maximum-likelihood scheme described here.

Molecular clock: The method described here assumes that the rate of mutation per unit time is the same on all branches. This is likely a reasonable assumption for the data considered here. Noncoding regions are less likely to be affected by natural selection than are the coding regions analyzed in other studies. Also, there is no reason to assume substantial differences in mutation rates (per generation) between humans and great apes. The data on generation times are sparse; EYRE-WALKER and KEIGHTLEY (1999) cite a time of $g = 23$ years in chimpanzees, while estimates of current human generation times are ~30 years (SIGUROARDÓTTIR *et al.* 2000; TREMBLAY and VÉZINA 2000). Average human generation times (over the last several million years) may be substantially smaller. Indeed, the CHEN and LI (2001) data show no evidence for more mutations on the chimpanzee branch than on the human branch (CHEN and LI 2001; results not shown), suggesting that the long-term average generation times for humans and chimpanzees are quite similar.

For other taxa, the clock assumption may not be appropriate. It would be straightforward to generalize the model to have different rates of evolution on different branches and to estimate these as well as species divergence times and ancestral population sizes. More sequence data and more computational time would be required to accurately estimate the additional parameters, and the method (with variable rates) may not be feasible with more than three species.

Nuisance parameters: Although the goal of this article is to estimate ancestral population sizes and species divergence times, the model presented here also includes other parameters, such as θ, ρ, or N_c/N_h. The reason ρ is included is not to estimate the recombination rate from divergence data (which would be somewhat challenging). Rather, the values of parameters like ρ affect the likelihoods, so some assumptions must be made about them. In the interest of computational tractability, I have chosen plausible values for θ, ρ, N_c/N_h, N_g/N_h, and N_o/N_h. Comparing model 1 with models 4 and 5 suggests that the choice of particular values for these other parameters may not affect the estimates of the parameters of interest. Further simulations show that this is true for a wider range of values ($\rho = 0.0005$–0.003/bp; $N_h \leq N_c, N_g, N_o \leq 6N_h$), although it should be pointed out that assuming no recombination (as previous methods do) leads to a likelihood of 0, due to the presence of several incompatibilities within loci. So, the estimates of N_a/N_h, T_1, T_2, and T_3 are robust to the assumptions made about the other parameters.

Speciation model: Estimates of N_a/N_h and T_1 provide information about the mean and the variance of the distribution of coalescent times of a single human and a single chimpanzee sequence. Under the simple speciation model considered here, greater variances in coalescent times must be the result of larger ancestral population sizes. Some researchers have suggested that there is often gene flow between "incipient species" (*e.g.*, WU 2001). A model of limited gene flow (prior to strict isolation) will lead to a greater variance in coalescent times. In particular, if there were gene flow after the initial divergence of the human and chimpanzee lines,

then the ancestral population size (before the initial divergence) would be overestimated. Further work must be done to develop methods that can distinguish between limited gene flow and large ancestral population sizes using orthologous sequence data.

I thank M. Hare, M. Przeworski, N. Takahata, J. Wakeley, and an anonymous reviewer for comments on an earlier version of this manuscript. J.D.W. was supported in part by a National Science Foundation Postdoctoral Fellowship in Bioinformatics.

LITERATURE CITED

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

Caballero, A., 1994 Developments in the prediction of effective population size. Heredity **73:** 657–679.

Chen, F.-C., and W.-H. Li, 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.

Deinard, A., and K. Kidd, 1999 Evolution of a *HOXB6* intergenic region within the great apes and humans. J. Hum. Evol. **36:** 687–703.

Easteal, S., and G. Herbert, 1997 Molecular evidence from the nuclear genome for the time frame of human evolution. J. Mol. Evol. **44:** S121–S132.

Ebersberger, I., D. Metzler, C. Schwarz and S. Pääbo, 2001 Genomewide comparison of DNA sequences between humans and chimpanzees. Am. J. Hum. Genet. **70:** 1490–1497.

Eyre-Walker, A., and P. D. Keightley, 1999 High genomic deleterious mutation rates in hominids. Nature **397:** 344–347.

Fearnhead, P., and P. Donnelly, 2002 Approximate likelihood methods for estimating local recombination rates. J. R. Stat. Soc. B **64:** 657–680.

Frisse, L., R. R. Hudson, A. Bartoszewicz, J. D. Wall, J. Donfack *et al.*, 2001 Gene conversion and different population histories may explain the contrast between polymorphism and linkage disequilibrium levels. Am. J. Hum. Genet. **69:** 831–843.

Gabunia, L., and A. Vekua, 1995 A Plio-Pleistocene hominid from Dmanisi, East Georgia, Caucasus. Nature **373:** 509–512.

Giannelli, F., T. Anagnostopolous and P. M. Green, 1999 Mutation rates in humans. II. Sporadic mutation-specific rates and rate of detrimental human mutations inferred from hemophilia B. Am. J. Hum. Genet. **65:** 1580–1587.

Haile-Selassie, Y., 2001 Late Miocene hominids from the Middle Awash, Ethiopia. Nature **412:** 178–181.

Harada, K., S. Kusakabe, T. Yamazaki and T. Mukai, 1993 Spontaneous mutation rates in null and band-morph mutations of enzyme loci in *Drosophila melanogaster*. Jpn. J. Genet. **68:** 605–616.

Harding, R. M., S. M. Fullerton, R. C. Griffiths, J. Bond, M. J. Cox *et al.*, 1997 Archaic African *and* Asian lineages in the genetic ancestry of modern humans. Am. J. Hum. Genet. **60:** 772–789.

Hey, J., 1994 Bridging phylogenetics and population genetics with gene tree models, pp. 435–449 in *Molecular Ecology and Evolution: Approaches and Applications*, edited by B. Schierwater, B. Streit, G. P. Wagner and R. Desalle. Birkhäuser Verlag, Basel, Switzerland.

Horai, S., K. Hayasaka, R. Kondo, K. Tsugane and N. Takahata, 1995 Recent African origin of modern humans revealed by complete sequences of hominoid mitochondrial DNAs. Proc. Natl. Acad. Sci. USA **92:** 532–536.

Hudson, R. R., 1983 Properties of a neutral allele model with intragenic recombination. Theor. Popul. Biol. **23:** 183–201.

Hudson, R. R., 1992 Gene trees, species trees and the segregation of ancestral alleles. Genetics **131:** 509–512.

Jones, P. A., W. M. Rideout, J. C. Shen, C. H. Spruck and Y. C. Tsai, 1992 Methylation, mutation and cancer. Bioessays **14:** 33–36.

Kaessmann, H., V. Wiebe and S. Pääbo, 1999 Extensive nuclear DNA sequence diversity among chimpanzees. Science **286:** 1159–1161.

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

Kimura, M., 1983 *The Neutral Theory of Molecular Evolution*. Cambridge University Press, Cambridge, UK.

Krawczak, M., and D. N. Cooper, 1991 Gene deletions causing human genetic disease: mechanisms of mutagenesis and the role of the local DNA sequence environment. Hum. Genet. **86:** 425–441.

Kreitman, M., 1983 Nucleotide polymorphism at the alcohol dehydrogenase locus of *Drosophila melanogaster*. Nature **304:** 412–417.

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

Kumar, S., and S. Subramanian, 2002 Mutation rates in mammalian genomes. Proc. Natl. Acad. Sci. USA **99:** 803–808.

Leakey, M. G., C. S. Feibel, I. Mcdougall and A. Walker, 1995 New four-million-year-old hominid species from Kanapoi and Allia Bay, Kenya. Nature **376:** 565–571.

Nachman, M. W., and S. L. Crowell, 2000 Estimate of the mutation rate per nucleotide in humans. Genetics **156:** 297–304.

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

Nielsen, R., and J. Wakeley, 2001 Distinguishing migration from isolation: a Markov chain Monte Carlo approach. Genetics **158:** 885–896.

Nordborg, M., 2001 Coalescent theory, pp. 179–212 in *Handbook of Statistical Genetics*, edited by D. Balding, M. Bishop and C. Cannings. Wiley, Chichester, UK.

Przeworski, M., and J. D. Wall, 2001 Why is there so little intragenic linkage disequilibrium in humans? Genet. Res. **77:** 143–151.

Przeworski, M., R. R. Hudson and A. Di Rienzo, 2000 Adjusting the focus on human variation. Trends Genet. **16:** 296–302.

Ruvolo, M., 1997 Molecular phylogeny of the hominoids: inferences from multiple independent DNA sequence data sets. Mol. Biol. Evol. **14:** 248–265.

Satta, Y., C. O'hUigin, N. Takahata and J. Klein, 1993 The synonymous substitution rate of the major histocompatibility complex loci in primates. Proc. Natl. Acad. Sci. USA **90:** 7480–7484.

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

Siguroardóttir, S., A. Helgason, J. R. Gulcher, K. Stefansson and P. Donnelly, 2000 The mutation rate in the human mtDNA control region. Am. J. Hum. Genet. **66:** 1599–1609.

Swisher, C. C., G. H. Curtis, T. Jacob, A. G. Getty, A. Suprijo *et al.*, 1994 Age of the earliest known hominids in Java, Indonesia. Science **263:** 1118–1121.

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., 1991 Trans-species polymorphism of *HLA* molecules, founder principle, and human evolution, pp. 29–49 in *Molecular Evolution of the Major Histocompatibility Complex*, edited by J. Klein and D. Klein. Springer, Heidelberg, Germany.

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

Takahata, N., 2001 Molecular phylogeny and demographic history of humans, pp. 299–305 in *Humanity From African Naissance to Coming Millennia—Colloquia in Human Biology and Palaeoanthropology*, edited by P. V. Tobias, M. A. Raath, J. Moggi-Cecchi and G. A. Doyle. Firenze University Press, Firenze, Italy.

Takahata, N., and Y. Satta, 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., and Y. Satta, 2002 Pre-speciation coalescence and the effective size of ancestral populations, pp. 52–71 in *Modern Developments in Theoretical Population Genetics*, edited by M. Slatkin and M. Veuille. Oxford University Press, Oxford.

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

Templeton, A. R., A. G. Clark, K. M. Weiss, D. A. Nickerson, E. Boerwinkle *et al.*, 2000 Recombinational and mutational hotspots within the human lipoprotein lipase gene. Am. J. Hum. Genet. **66:** 69–83.

Tremblay, M., and H. Vézina, 2000 New estimates of intergenera-

tional time intervals for the calculation of age and origins of mutations. Am. J. Hum. Genet. **66:** 651–658.

Wakeley, J., and J. Hey, 1997 Estimating ancestral population parameters. Genetics **145:** 847–855.

Wall, J. D., 2000 A comparison of estimators of the population recombination rate. Mol. Biol. Evol. **17:** 156–163.

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

Weiss, G., and A. Von Haeseler, 1998 Inference of population history using a likelihood approach. Genetics **149:** 1539–1546.

White, T. D., G. Suwa and B. Asfaw, 1994 *Australopithecus ramidus*, a new species of early hominid from Aramis, Ethiopia. Nature **371:** 306–312.

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

Wu, C.-I, 2001 The genic view of the process of speciation. J. Evol. Biol. **14:** 851–866.

Yang, Z., 1997 On the estimation of ancestral population sizes of modern humans. 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.

Yu, A., C. Zhao, Y. Fan, W. Jang, A. J. Mungall *et al.*, 2001 Comparison of human genetic and sequence-based physical maps. Nature **409:** 951–953.

Communicating editor: N. Takahata