GBE

A Likelihood Ratio Test of Speciation with Gene Flow Using Genomic Sequence Data

Ziheng Yang*

Galton Laboratory, Department of Biology, University College London, United Kingdom and School of Life Sciences, Sun Yat-sen University, Guangzhou, China

*Corresponding author: E-mail: z.yang@ucl.ac.uk.

Accepted: 14 March 2010

Abstract

Genomic sequence data may be used to test hypotheses about the process of species formation. In this paper, I implement a likelihood ratio test of variable species divergence times over the genome, which may be considered a test of the null model of allopatric speciation without gene flow against the alternative model of parapatric speciation with gene flow. Two models are implemented in the likelihood framework, which accommodate coalescent events in the ancestral populations in a phylogeny of three species. One model assumes a constant species divergence time over the genome, whereas another allows it to vary. Computer simulation shows that the test has acceptable false positive rate but to achieve reasonable power, hundreds or even thousands of genomic loci may be necessary. The test is applied to genomic data from the human, chimpanzee, and gorilla.

Key words: population size, coalescent, maximum likelihood, speciation, gene flow, parapatric speciation, allopatric speciation.

Introduction

Genomic sequence data provide information not only about population demographic processes of modern species (Wilson et al. 2003; Heled and Drummond 2008) but also about such processes in extinct ancestral species (Rannala and Yang 2003) and even about the mode and timing of the speciation process itself. Takahata (1986) pointed out that sequences from multiple genomic regions of two closely related extant species can be used to estimate the population size of their common ancestor, relying on the fact that the coalescent time in the ancestral population fluctuates over loci at random, in proportion to the ancestral population size. The sequence distance between two species is comprised of two parts, due to the evolution since the time of species separation (τ) and to the evolution during the coalescent time t in the common ancestor. Although τ is constant over the whole genome, t varies over genomic regions according to the exponential distribution with both the mean and the standard deviation (SD) equal to $2N$ generations, where N is the effective population size of the ancestor. Takahata et al. (1995) extended this analysis to three species, using maximum likelihood to account for uncertainties in the gene tree topology and coalescent times. The past few years have seen considerable improvements in the statistical methodology for analyzing multiple-species multiple-loci data sets, particularly concerning reconstruction of species phylogenies in presence of gene tree conflicts (for reviews, see Rannala and Yang 2008; Liu et al. 2009).

Genomic data may also shed light on the mode and timing of the process of species formation (Patterson et al. 2006; Burgess and Yang 2008). Wu and Ting (2004) argue that while the species divergence time τ may be constant over genomic regions if speciation is allopatric, with gene flow ceasing immediately at the time of species separation, τ should vary if speciation is parapatric and reproductive isolation develops gradually over a period of time. Osada and Wu (2005; see also Zhou et al. 2007) explored this idea to develop a likelihood ratio test (LRT) of the null hypothesis that τ is constant between two kinds of loci against the alternative that τ is variable. With only two species in the comparison, the test may have low power and may be very sensitive to variable mutation rates among loci. The information about variable τs over loci comes mostly from the variation, among loci, in sequence divergence between the two species. However, a large variation in sequence divergence can be explained by any of the following reasons: variable

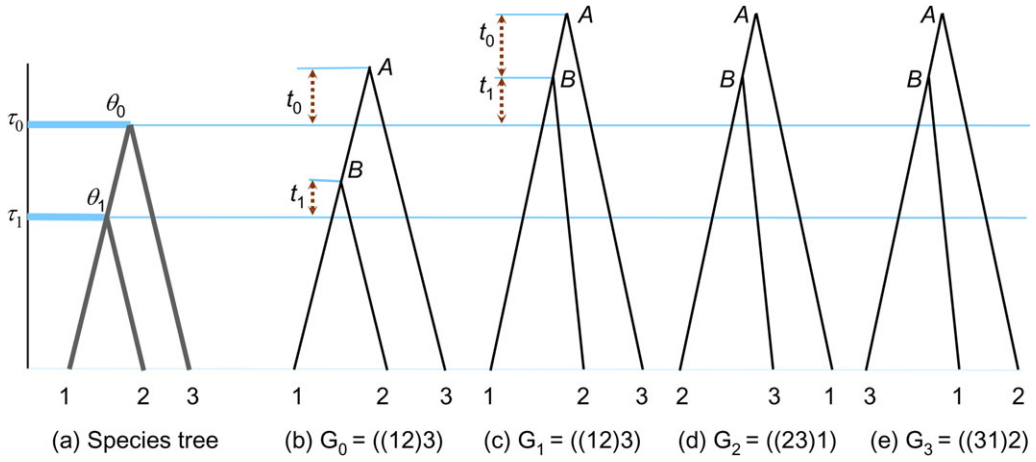

Fig. 1.—(*a*) The species tree ((12)3) for three species, showing the parameters in model M0: θ_0, θ_1, τ_0, and τ_1. The four possible gene trees for any locus are shown in *b*–*e*. If sequences *a* and *b* coalesce in the common ancestor of species 1 and 2, the resulting gene tree will be G_0 (*b*). Otherwise three gene trees G_1, G_2, and G_3 are possible as shown in (*c*)–(*e*).

mutation rates, a large ancestral population size, and variable species divergence times. The simple model of speciation without gene flow with a large ancestral θ may explain the sequence data nearly as well as the more complex model of speciation with gene flow, so that the test will likely lack power.

The problem may be alleviated somewhat by inclusion of a close outgroup species. With three species (fig. 1), the gene tree can differ from the species tree, and such conflicts between the gene tree and the species tree provide information about the ancestral population size. The species-tree gene-tree mismatch probability is $\frac{2}{3}e^{-2(\tau_0-\tau_1)/\theta_1}$, or $\frac{2}{3}$ the probability that the sequences from species 1 and 2 do not coalesce in the common ancestor of species 1 and 2 (fig. 1) (Hudson 1983). Furthermore, the outgroup species may provide information about the relative mutation rate at the locus, so that the test may become less sensitive to mutation rate variation. For example, a large between-species distance d_{12} can be due to a long coalescent time in the ancestor or a high mutation rate at the locus, but if d_{23} and d_{31} are small at the locus, the former explanation becomes more likely. Of course, the gene tree topology and branch lengths involve substantial uncertainties due to lack of information in the alignment at each locus, but such uncertainties can be dealt with properly in a standard likelihood approach. Indeed, Yang (2002) implemented a maximum likelihood method for the case of three species under the simple allopatric speciation model (fig. 1). The JC model (Jukes and Cantor 1969) was used to correct for multiple hits. This is an extension of the maximum likelihood method of Takahata et al. (1995), which assumes the infinite sites mutation model. The likelihood calculation involves 2D integrals, which were calculated using Mathematica.

In this paper, I improve the computational algorithm of Yang (2002) so that it can be used for larger data sets with more loci. Numerical integration using Mathematica is slow, so I use Gaussian quadrature method instead. I then implement a new model that allows the species divergence time to vary among loci at random. The new model is compared with the old model to formulate an LRT of constant species divergence time τ_1 (fig. 1). This may be interpreted as a test of the null model of speciation without gene flow against the alternative model of speciation with gene flow. Although gene flow at the early stages of allopatric speciation is imaginable, parapatric and sympatric speciation appears to be the more natural scenario of speciation with gene flow. Thus, the test may also be considered a test of the null model of allopatric speciation against the alternative model of parapatric (and sympatric) speciation. Computer simulations are conducted to assess the sampling errors in parameter estimates and to examine the false positive rate and power of the test. The method is then applied to a data set of genomic sequences from the human, chimpanzee, and gorilla (Burgess and Yang 2008).

Theory

The Model of Constant Speciation Time (Model M0)

I briefly describe the model of Yang (2002) to introduce the notation and to discuss the computational issues involved. The species tree ((1, 2), 3) is assumed known (fig. 1a), and the two ancestral species are referred to as 12 and 123. There are four parameters in the model: $\theta_0 = 4N_0\mu$ for the ancestor 123, $\theta_1 = 4N_1\mu$ for the ancestor 12, and two species divergence times τ_0 and τ_1. Here, μ is the mutation rate, N_0 and N_1 are the two ancestral (effective) population sizes, whereas τ_0 and τ_1 are species divergence times multiplied by the mutation rate.

The data consist of DNA sequences from multiple neutral loci, with one sequence from each species at each locus. It is assumed that there is no recombination within a locus and free recombination between loci. Each population is assumed to be random mating, and there is no gene flow since species separation.

Under the Jukes and Cantor (1969) mutation model, the sequence alignments at any locus i can be summarized as the counts of sites, $D_i = \{n_{i0}, n_{i1}, n_{i2}, n_{i3}, n_{i4}\}$, for five site patterns *xxx*, *xxy*, *yxx*, *xyx*, and *xyz*, where *x*, *y*, and *z* are any different nucleotides. Sites with ambiguities and alignment gaps are removed. We define branch lengths b_0 and b_1 as the lengths of branches AB and B1, respectively, in gene tree G_1 (fig. 1). Branch lengths in other gene trees are defined similarly. Given the gene tree G_1 (or G_0) and branch lengths b_0 and b_1, the probabilities of observing the five site patterns are

$$\begin{aligned} p_0 &= \tfrac{1}{16}(1 + 3e^{-8b_1/3} + 6e^{-8(b_0+b_1)/3} + 6e^{-(8b_0+12b_1)/3}),\\ p_1 &= \tfrac{1}{16}(3 + 9e^{-8b_1/3} - 6e^{-8(b_0+b_1)/3} - 6e^{-(8b_0+12b_1)/3}),\\ p_2 &= \tfrac{1}{16}(3 - 3e^{-8b_1/3} + 6e^{-8(b_0+b_1)/3} - 6e^{-(8b_0+12b_1)/3}),\\ p_3 &= p_2,\\ p_4 &= \tfrac{1}{16}(6 - 6e^{-8b_1/3} - 12e^{-8(b_0+b_1)/3} + 12e^{-(8b_0+12b_1)/3}) \end{aligned} \quad (1)$$

(Yang 1994). The conditional probabilities of data at locus i given the gene tree and branch lengths are given by the multinomial distribution as

$$\begin{aligned} P(D_i|G_1, b_0, b_1) &= p_0^{n_{i0}} p_1^{n_{i1}} p_2^{n_{i2}+n_{i3}} p_4^{n_{i4}},\\ P(D_i|G_2, b_0, b_1) &= p_0^{n_{i0}} p_1^{n_{i2}} p_2^{n_{i3}+n_{i1}} p_4^{n_{i4}},\\ P(D_i|G_3, b_0, b_1) &= p_0^{n_{i0}} p_1^{n_{i3}} p_2^{n_{i1}+n_{i2}} p_4^{n_{i4}}. \end{aligned} \quad (2)$$

The unconditional probability of data D_i at locus i is an average over the gene trees and branch lengths (i.e., over coalescent times t_0 and t_1)

$$\begin{aligned} &f(D_i|\theta_0, \theta_1, \tau_0, \tau_1)\\ &= \int_0^\infty \int_0^{2(\tau_0-\tau_1)/\theta_1} P(D_i|G_0, \tau_0 - \tau_1 - \tfrac{1}{2}\theta_1 t_1 + \tfrac{1}{2}\theta_0 t_0, \tau_1\\ &+ \tfrac{1}{2}\theta_1 t_1) \times e^{-t_1} e^{-t_0} dt_1 dt_0\\ &+ e^{-2(\tau_0-\tau_1)/\theta_1} \int_0^\infty \int_0^\infty \Big[\sum_{k=1}^{3} P(D_i|G_k, \tfrac{1}{2}\theta_0 t_0, \tau_0 + \tfrac{1}{2}\theta_0 t_1)\Big]\\ &e^{-3t_1} e^{-t_0} dt_1 dt_0 \end{aligned} \quad (3)$$

(Yang 2002: eq. 8). The first term in the equation corresponds to gene tree G_0 and the second to the three gene trees G_1, G_2, and G_3 (fig. 1). Note that with time measured in $2N$ generations, the coalescent time t has an exponential distribution (with mean 1 for two lineages or mean 1/3 for three lineages) and contributes a mutational distance of $\frac{1}{2}\theta t$.

Finally, the likelihood is a product over all the L loci

$$f(D|\theta_0, \theta_1, \tau_0, \tau_1) = \prod_{i=1}^{L} f(D_i|\theta_0, \theta_1, \tau_0, \tau_1). \quad (4)$$

Parameters θ_0, θ_1, τ_0, and τ_1 are estimated by numerical maximization of the log likelihood $\ell = \log\{f(D|\theta_0, \theta_1, \tau_0, \tau_1)\}$. The numerical optimization routine used here (Yang 1997) deals with lower and upper bounds but not general linear inequality constraints such as $\tau_1 < \tau_0$. Thus, the transformation $x_1 = \tau_1/\tau_0$ is used instead of τ_1, with $0 < x_1 < 1$.

Numerical Integration

Each evaluation of the likelihood function (4) requires calculation of $2L$ 2D integrals. Yang (2002) used Mathematica to calculate them numerically. This was found to be reliable but quite slow. In this paper, I apply Gaussian quadrature, using the Gauss-Legendre rule (e.g., Kincaid and Cheney 2002, p. 492–501), by which a 1-D integral is approximated using a sum of K terms

$$\int_a^b f(x)dx \approx \frac{b-a}{2}\sum_{i=1}^{K} w_i f\Big(\frac{b+a}{2} + \frac{b-a}{2}x_i\Big), \quad (5)$$

where the points x_i and weights w_i are given according to the Gauss-Legendre rule. Note that the number of points K is not a parameter in the model but affects the computation, with a larger K producing more accurate but computationally more expensive approximation. Two-dimensional integrals can be calculated by repeated use of this approximation, with the computation proportional to K^2. Tests using the data of Chen and Li (2001), which include $L = 53$ loci each of about 500 bp, in comparison with the results of Yang (2002), suggest that $K = 8$ or 16 provide adequate approximation under this model. Maximum likelihood iteration for the data set takes about 30 min using the old algorithm and ~5 s using the new one on the same PC.

The probabilities $P(D_i|G_k, b_0, b_1)$ of equation (2) are very small and vary over many orders of magnitude depending on b_0 and b_1. To avoid underflows and overflows, the highest log likelihood at the locus, ℓ_{max}, calculated at the maximum likelihood tree topology and branch lengths, is used for scaling: the integrands of equation (3) are divided by $e^{\ell_{max}}$ before they are summed up (Yang 2006: eq. 9.9). Tests suggest that with this scaling, the algorithm is feasible for up to 10 kb at each locus.

Model of Variable τ_1 among Loci (Model M1)

This model allows the divergence time τ_1 to vary among loci. Species 3 is considered an outgroup and its divergence time (τ_0) from the common ancestor of species 1 and 2 is assumed to be constant. No theory appears to exist to predict how τ_1 should vary among loci under a model of parapatric speciation with gene flow, so my choice here is somewhat

arbitrary (see Discussion). One may use the gamma distribution but the truncation (so that $\tau_1 < \tau_0$) makes it awkward to interpret the model parameters. The beta distribution appears to be quite flexible and is implemented here. The density is

$$f(\tau_1;\tau_0,p,q)=\frac{1}{B(p,q)}\left(\frac{\tau_1}{\tau_0}\right)^{p-1}\left(1-\frac{\tau_1}{\tau_0}\right)^{q-1}\cdot\frac{1}{\tau_0},\quad 0<\tau_1<\tau_0. \tag{6}$$

Here τ_0, p, and q are the parameters of the distribution. The model is equivalent to assuming that the transformed variable $x_1 = \tau_1/\tau_0$ has the familiar two-parameter beta distribution: $x_1 \sim$ beta(p, q) with $0 < x_1 < 1$. The distribution is uniform if $p = 1$ and $q = 1$, has a single mode if $p > 1$ and $q > 1$, and can take a variety of shapes depending on p and q. The mean of the distribution is $\bar{x}_1 = p/(p+q)$ and the variance is $s^2 = pq/[(p+q)^2(p+q+1)]$. For easy comparison with model M0, I use $\bar{x}_1$ and q instead of p and q as parameters of the model, with $p = \bar{x}_1/(1-\bar{x}_1)\cdot q$, $0 < \bar{x}_1 < 1$ and $0 < q < \infty$. Thus, model M1 involves five parameters: θ_0, θ_1, τ_0, $\bar{x}_1 = \bar{\tau}_1/\tau_0$, and q. With this formulation, parameter q is inversely related to the variance in τ_1, and the null model of constant τ_1 is represented by $q = \infty$.

The probability of data at a locus is then

$$f(D_i|\theta_0,\theta_1,\tau_0,\bar{\tau}_1,q)=\int_0^1 f(D_i|\theta_0,\theta_1,\tau_0,x_1\tau_0)f(x_1|\bar{x}_1,q)\mathrm{d}x_1, \tag{7}$$

where $f(D_i|\theta_0, \theta_1, \tau_0, x_1\tau_0)$ is given by equation (3) with $\tau_1 = x_1\tau_0$, $f(x_1|\bar{x}_1, q)$is the beta density. Under this model, the integrals are 3D, so that the computation involved in Gaussian quadrature is proportional to K^3.

To let the algorithm focus on the region where the integrand is large, the integral limits in equation (7) are changed to max(0, $\bar{x}_1 - 5s$) and min(1, $\bar{x}_1 + 5s$), where s is the SD of the beta distribution. For the same K, the approximation to the 3-D integrals under M1 is poorer than the approximation to the 2-D integrals under M0. Furthermore, the approximation is poorer for small qs than for large qs (fig. 2). Tests suggest that $K = 16$ provides adequate approximation: this value is used in the simulation and analysis in this paper.

The LRT

When $q = \infty$, model M1 reduces to the simple model of a constant τ_1. The two models are thus nested and can be compared using an LRT. Let the test statistic be $2\Delta\ell = 2(\ell_1 - \ell_0)$, where ℓ_0 and ℓ_1 are the log likelihood values under the two models. Because $q = \infty$ is at the boundary of the parameter space of model M1, the standard χ^2_1 approximation breaks down. Instead, the null distribution is the 50:50 mixture of point mass 0 and χ^2_1 (Self and Liang 1987). The critical values are 2.71 at 5% and 5.41 at 1% (as opposed to 3.84 for 5% and 6.63 for 1% for χ^2_1. The P value for the mixture is half the P value from χ^2_1 for the same test statistic.

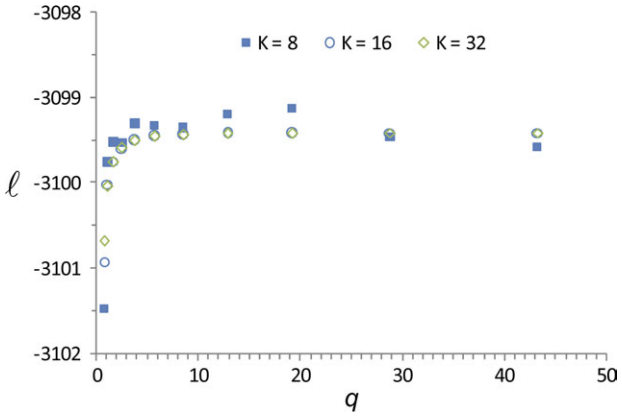

Fig. 2.—The approximate log likelihood under model M1 (parapatric speciation) for different values of q calculated using the Gauss-Legendre quadrature with K points. The data of Chen and Li (2001) are used. Parameters other than q are fixed at their estimates under model M0: $\theta_0 = 0.003057$, $\theta_1 = 0.000990$, $\tau_0 = 0.006283$, $\bar{\tau}_1 = 0.005194$ (or $x_1 = 0.8267$) (Yang 2002). The values for $K = 16$ and 32 are indistinguishable for $q > 0.75$. The MLE of q appears to be ∞. The log likelihood at $q = \infty$ (i.e., model M0) is −3099.41, whereas the approximate values at $q = 50$ are –3099.60 for $K = 8$ and –3099.41 for $K \geq 16$.

Mutation Rate Variation among Loci

The information concerning ancestral θs and possible variation in divergence time τ comes mostly from the variation in the gene tree topology and branch lengths among loci. As different mutation rates can cause such variation as well, rate variation among loci may be a serious concern. Although rates may be nearly constant among neutral loci (such as the hominoid genomic data analyzed later in this paper), they may vary considerably over functional regions or protein-coding genes. Because different genes are under different selective constraints, they have different proportions of neutral mutations and different neutral mutation rates.

Following Yang (2002), an outgroup species may be used to estimate the relative rates for the loci, which may be used as constants in the likelihood calculation. If the rate for locus i is r_i, the branch lengths in equations (2) and (3) are simply multiplied by r_i. As the relative rates are scaled to have mean 1, parameters (θs and τs) are all defined using the average rate across all loci.

Results

Analysis of Simulated Data

Three simulations are conducted to examine the sampling errors of the maximum likelihood estimates (MLEs) and the type-I and type-II errors of the LRT. The first simulates data under model M0 to examine the sampling errors in

Table 1
Maximum Likelihood Estimates (Mean ± SD) of Parameters under Model M0

Parameters	θ_0	θ_1	τ_0	τ_1
Hominoid set	(0.005)	(0.005)	(0.006)	(0.004)
$L = 10$[a]	0.0040 ± 0.0028	0.0083 ± 0.0129	0.0065 ± 0.0013	0.0041 ± 0.0018
$L = 100$	0.0049 ± 0.0009	0.0055 ± 0.0040	0.0060 ± 0.0004	0.0040 ± 0.0008
$L = 1{,}000$	0.0050 ± 0.0003	0.0051 ± 0.0011	0.0060 ± 0.0001	0.0040 ± 0.0002
Mangrove set	(0.01)	(0.01)	(0.02)	(0.01)
$L = 10$	0.0082 ± 0.0059	0.0099 ± 0.0083	0.0209 ± 0.0027	0.0106 ± 0.0026
$L = 100$	0.0099 ± 0.0017	0.0101 ± 0.0021	0.0201 ± 0.0008	0.0100 ± 0.0007
$L = 1{,}000$	0.0100 ± 0.0005	0.0100 ± 0.0007	0.0200 ± 0.0002	0.0100 ± 0.0002

Note.—The true parameter values are shown in the parentheses.
[a] In 4.7% of replicates, $\hat{\theta}_1$ is ∞, and those estimates are not used in calculation of the means and SDs.

the MLEs of model parameters. Two sets of parameter values are used in the simulation, roughly based on estimates from the hominoids: $\theta_0 = 0.005$, $\theta_1 = 0.005$, $\tau_0 = 0.006$, $\tau_1 = 0.004$ (Burgess and Yang 2008) and from the mangroves: $\theta_0 = 0.01$, $\theta_1 = 0.01$, $\tau_0 = 0.02$, $\tau_1 = 0.01$ (Zhou et al. 2007). The JC69 mutation model, with constant rate among loci, is used both to simulate and to analyze the data. Given the parameter values, the probabilities of the five site patterns are calculated using equation (1) and the counts of sites at each locus (n_{i0}, n_{i1}, n_{i2}, n_{i3}, n_{i4}) are generated by sampling from the multinomial distribution. Each locus has 500 sites. Each replicate data set consists of L loci, which are analyzed to obtain the MLEs of the parameters under model M0. The number of replicates is 1,000.

The means and SDs of the parameter estimates under model M0 are listed in table 1. For the hominoid parameter set, estimates of θ_0 and θ_1 are quite poor with $L = 10$ loci, although τ_1 is well estimated. Estimates of θ_1 have a positive bias. The fact that θ_1 is more poorly estimated than θ_0 may seem counterintuitive as one might expect it to be easier to estimate parameters for recent ancestors (such as θ_1) than for ancient ancestors (such as θ_0). Nevertheless, this expectation may not be correct. For the hominoid parameter set, the two speciation times are close, so that there was little chance for coalescent events to occur during that time interval, which would provide information about θ_1. With 100 or 1,000 loci, all parameters are well estimated.

For the mangrove set, the parameters are greater so that the sequences are more informative. Indeed, even with $L = 10$ loci, all parameters except θ_0 are well estimated. The difference in the overall performance of the method between the two parameter sets appears to be mainly due to the different mutation rates (i.e., larger values of θ and τ for the mangrove set). The more accurate estimation of θ_1 for the mangrove set may also be due to the larger time interval between the two speciation events and thus more chances for coalescent events during that time interval: the probability of gene tree G_0 is $1 - e^{-2(\tau_0-\tau_1)/\theta_1} = 0.55$ for the mangrove set and 0.86 for the hominoid set. For both sets, the results are consistent with the expectation that a 10-fold increase in the number of loci leads to $\sqrt{10}$-fold reduction in the SD.

The second simulation examines the type-I error rate of the LRT implemented in this paper. Data are simulated under model M0 using the two sets of parameter values (for hominoids and mangroves). Each locus has 500 bp. The number of replicates is 200. Each replicate data set is analyzed using models 0 and 1 to calculate the test statistic $2\Delta\ell = 2(\ell_1 - \ell_0)$. The results are shown in table 2, with the significance level set at 5%. The test appears to be conservative, with the false positive rate <5%, when the data contain little information (i.e., when $L = 10$ or 100 for the hominoid set and when $L = 10$ for the mangrove set). With more loci or with a higher mutation rate, the false positive rate becomes close to the nominal 5%.

The third simulation examines the power of the LRT. Data are simulated under model M1, using $q = 1.2$ (which is the estimate from the hominoid data; see below). As before, two sets of parameter values for θ_0, θ_1, τ_0, and τ_1 are used. Again each locus has 500 sites, and the number of replicates is 200. The results are shown in table 2. For the hominoid set, the test has virtually no power (<5%) with $L = 10$ or 100 loci and moderate power (52%) when $L = 1{,}000$. For the mangrove set, the power is quite high (78%) with 100 loci and reaches 100% when $L = 1{,}000$. The large difference between the two parameter sets lies mainly in the near 2-fold difference in mutation rate and the information content in the sequence data. Longer sequences in each alignment are expected to improve the power just like a higher mutation rate (Felsenstein 2005), but this effect is not evaluated here.

Analysis of Hominoid Data

Here, I apply the LRT to the genomic sequences of the human, chimpanzee, and gorilla from Burgess and Yang (2008). These data are an updated version of the data of Patterson et al. (2006), updated and recurated by Burgess and Yang (2008) to incorporate more recent genome assembly sequences and to generate high-quality alignments of genomic regions instead of single variable sites. Filters

Table 2
False Positive Rate and Power of the LRT in Simulations

Simulation Model	L = 10	100	1,000
False Positive Rate			
M0, $q = \infty$ (hominoid)	0.00	0.01	0.07
M0, $q = \infty$ (mangroves)	0.00	0.06	0.05
Power			
M1, q = 1.2 (hominoid)	0.00	0.02	0.52
M1, q = 1.2 (mangroves)	0.06	0.78	1.00

NOTE.—Proportion of simulated replicates in which the test statistic exceeds 2.71, the critical value at the 5% level.

were applied to remove the error-prone ends of whole-genome shotgun reads, as well as coding regions, repeats, RNA genes, and low-complexity regions. As the model assumes free recombination between loci and no recombination within locus, the data were filtered so that each locus (genomic region) was at least 1 kb away from known genes, and every two loci had a minimum separation of 10 kb. The resulting "neutral" data set comprised 14,663 autosomal loci and 783 X-linked loci for five species: human (H), chimpanzee (C), gorilla (G), orangutan (O), and macaque (M). The mean locus length was 508 bp. The likelihood method of this paper can analyze three species only, so the human, chimpanzee, and gorilla sequences are used. To test the impact of mutation rate variation among loci, the orangutan sequence is used as the outgroup to calculate relative mutation rates for the loci (Yang 2002). Thus, some loci at which the orangutan sequence is missing are excluded in the analysis, leaving 9,861 autosomal loci and 510 X loci. The data for the 22 human autosomal chromosomes are analyzed separately and are then combined in one analysis. Sites with alignment gaps and ambiguity nucleotides are removed. Although Burgess and Yang (2008) modeled sequencing errors and violations of the molecular clock, those factors were found to have only minor impact on estimation of parameters concerning the human, chimpanzee, and gorilla in the analysis of the curated data (compare tables 2 and 5 in Burgess and Yang 2008). In this paper, sequencing errors are ignored and the molecular clock is assumed.

The results of the LRT are shown in table 3. When the mutation rate is assumed to be constant among loci, the test is significant at 3 out of the 22 autosomes. Using the relative rates calculated from comparison with the orangutan, the test is significant at 6 out of the 22 autosomes, as well as for the X chromosome. If the apparent variation in τ_{HC} among loci is due to mutation rate variation, accounting for variable mutation rates among loci should lead to a reduction in the number of significant results. Thus, there seems to be little evidence for variable rates among loci in those data (for the similarity of parameter estimates under the basic model and the variable-rates models, see also Burgess and Yang 2008; table 2), and the LRT is not misled by possible rate variation among loci in this analysis. The average rates for the 22 autosomes, as indicated by the average JC69 distance between HCG and the orangutan, are very homogeneous (table 3), indicating little rate differences among the chromosomes.

When all the 9,861 autosomal loci are used in the same analysis, the LRT is highly significant whether the mutation rate is assumed to be constant or variable across loci. There is thus evidence for variable τ_1 over the genome. This model, although not so extreme as the large-scale hybridization model envisaged by Patterson et al. (2006), is incompatible with the simple model of a constant τ over the genome. The evidence should perhaps not be considered overwhelming, given the huge number of loci used in the test. It has been suggested that the LRT tends to reject the null model too often in large data sets and that the Bayesian method may provide a more accurate assessment of the evidence in the data concerning the models (e.g., Schwarz 1978). A Bayesian implementation of the same test would make it possible to compare the different methodologies using the same data.

Maximum likelihood estimates of the parameters under model M0 obtained from the analysis of all the autosomal loci are as follows: $\hat{\theta}_{HCG}$ = 0.00358 ± 0.00008, $\hat{\theta}_{HC}$ = 0.00431 ± 0.00025, $\hat{\tau}_{HCG}$ = 0.00661 ± 0.00004, $\hat{\tau}_{HC}$ = 0.00432 ± 0.00007. The standard errors (SEs) are very small due to the large size of the data. The estimates of τ_{HC} and τ_{HCG} are very similar to those of Burgess and Yang (2008), although those of θ_{HC} and θ_{HCG} are more different (table 3). I analyzed the same data using the Bayesian program of Rannala and Yang (2003), using the gamma prior $G(2, 2{,}000)$ for all θs and $\tau_{HCG} \sim G(2, 300)$ with mean 0.0067. The means and SDs of the posterior distribution are $\hat{\theta}_{HCG}$ = 0.00360 ± 0.00008, $\hat{\theta}_{HC}$ = 0.00419 ± 0.00027, $\hat{\tau}_{HCG}$ = 0.00660 ± 0.00004, $\hat{\tau}_{HC}$ = 0.00435 ± 0.00008. These are virtually identical to the MLEs and SEs obtained in the likelihood analysis. Further tests (supplementary table S1, Supplementary Material online) suggest that the differences between the MLEs of this paper and the Bayesian estimates of Burgess and Yang (2008) are not due to different estimation methods or to removal of some loci or of sites with ambiguous nucleotides: instead they are due to exclusion of orangutan and macaque in the present data set. The posterior mean of θ_{HC} is about 0.0042 in the HCG data sets but 0.0060 in the HCGO data sets and 0.0064 in the HCGOM data sets. The reasons for those differences are unclear. The molecular clock assumption is most likely violated when the macaque is included in the analysis. However, accommodating the higher rate in the macaque lineages was found to have very minor impact on estimates of θ_{HC}, clearly insufficient to explain the differences observed here (Burgess and Yang 2008: table 2e). Estimates of the other parameters are all very similar in the different data sets and analyses, with the posterior means

Table 3
Maximum Likelihood Estimates of Parameters under Model M0 and the LRT Statistic for Hominoid Genomic Loci from Each Chromosome

Chromosome	L	$d_{HCG\text{-}O}$	Constant Rate among Loci					Variable Rates among Loci				
			θ_{HCG}	θ_{HC}	τ_{HCG}	τ_{HC}	$2\Delta\ell$	θ_{HCG}	θ_{HC}	τ_{HCG}	τ_{HC}	$2\Delta\ell$
1	759	0.0346	3.60	3.34	6.44	4.20	0.00	3.65	3.86	6.52	4.09	0.40
2	1009	0.0351	3.34	5.33	6.75	4.20	1.01	3.32	5.43	6.86	4.28	0.64
3	732	0.0358	3.77	3.65	6.25	4.44	0.03	3.53	4.44	6.47	4.34	1.92
4	768	0.0351	3.78	4.37	6.68	4.38	2.47	3.94	5.92	6.72	4.01	5.27
5	788	0.0351	3.46	3.50	6.45	4.45	2.05	3.50	4.40	6.51	4.23	0.11
6	627	0.0342	3.18	4.53	6.30	3.92	3.79	3.80	4.61	6.12	3.89	10.12
7	506	0.0346	4.11	3.13	6.47	4.56	2.62	4.13	3.48	6.55	4.48	3.69
8	623	0.0364	4.12	3.83	6.57	4.57	4.73	3.42	5.02	6.95	4.38	4.43
9	381	0.0332	3.69	5.39	6.93	4.28	1.98	3.62	5.93	7.06	4.22	0.50
10	458	0.0357	2.89	5.71	6.86	3.81	2.73	3.53	4.76	6.69	4.10	1.96
11	427	0.0348	3.56	8.90	6.49	3.52	0.02	3.82	9.69	6.47	3.44	1.40
12	468	0.0347	2.91	3.17	6.66	4.53	2.53	3.06	3.04	6.64	4.59	0.26
13	431	0.0356	3.44	4.17	6.68	4.34	1.83	3.93	4.43	6.54	4.24	2.80
14	325	0.0345	3.43	5.46	6.36	3.80	2.04	2.95	7.04	6.72	3.59	3.33
15	277	0.0352	2.82	7.45	7.28	3.94	0.17	3.05	6.97	7.21	4.12	0.29
16	254	0.0382	4.81	5.58	7.05	4.59	0.00	4.99	5.47	7.06	4.72	1.20
17	202	0.0350	3.71	0.96	6.39	5.36	0.09	3.33	0.85	6.62	5.50	1.50
18	327	0.0359	3.60	4.40	6.57	4.58	1.12	3.56	4.69	6.75	4.64	0.04
19	84	0.0391	3.50	0.18	7.12	6.39	0.77	2.63	3.11	7.77	5.20	1.48
20	215	0.0364	2.87	5.13	7.38	4.26	1.53	3.73	3.10	7.05	4.90	1.06
21	122	0.0376	4.00	0.49	7.02	6.49	0.00	2.97	4.42	7.66	5.13	0.00
22	78	0.0399	2.89	4.84	8.23	4.75	1.95	2.88	5.17	8.25	4.68	0.16
A	9861	0.0353	3.58	4.31	6.61	4.32	36.92	3.63	4.77	6.68	4.26	46.08
X	510	0.0282	3.05	1.42	5.21	3.62	0.70	2.38	2.27	5.58	3.32	4.87
A(BY08)[a]	14,663		3.4	6.5	6.7	4.1			3.3	6.1	6.3	3.9
X(BY08)[a]	783		2.0	2.6	5.4	3.1						

NOTE.—θ and τ estimates are scaled by 10^3.
[a] The posterior means from Burgess and Yang (2008: table 2).

to be in the range 0.0033–0.0036 for θ_{HCG}, 0.0063–0.0068 for τ_{HCG}, and 0.0039–0.0043 for τ_{HC} (supplementary table S1, Supplementary Material online). Similar patterns are noted for the X chromosome loci (supplementary table S2, Supplementary Material online). Estimates of θ_{HC} was 0.0014–0.0016 in the HCG data sets but 0.0023 in HCGO and 0.0026 in HCGOM data sets. Inclusion of orangutan and macaque also caused the estimates of θ_{HCG} to become smaller, with the posterior means to be 0.0029–0.0030 in the HCG, 0.0024 in the HCGO, and 0.0020–0.0022 in the HCGOM data sets.

Table 4 shows the correlations between parameter estimates in the analysis of the hominoid autosomal loci. Estimates of θ_{HCG} and τ_{HCG} are strongly correlated, as are those of θ_{HC} and τ_{HC}. As in the simulated data sets (table 1), θ_{HCG} is more precisely estimated than θ_{HC}.

The estimates under model M1 from analysis of all the autosomal loci are $\hat{\theta}_{HCG} = 0.00367 \pm 0.00008$, $\hat{\theta}_{HC} = 0.00137 \pm 0.00014$, $\hat{\tau}_{HCG} = 0.00657 \pm 0.00004$, $\bar{\tau}_{HC} = 0.00530 \pm 0.00006$, and $\hat{q} = 1.189 \pm 0.068$ for the beta model of τ_{HC} variation. The estimated q is rather small, consistent with the rejection of the null model M0 by the LRT. The estimated beta distribution beta(5.0, 1.2) is shown in figure 3, which implies that for most genomic regions, τ_{HC} is near τ_{HCG} or if there were migrations at the time of separation of the human and chimpanzee, the gene flow had ceased a long time ago. As expected, estimates of q and θ_{HC} are strongly negatively correlated (table 4).

Table 4
Correlations of Parameter Estimates for the Hominoid Autosomal Loci (9,861 Loci)

	θ_{HCG}	θ_{HC}	τ_{HCG}	τ_{HC}
Model M0				
θ_{HCG}				
θ_{HC}	−0.34			
τ_{HCG}	**−0.71**	0.29		
τ_{HC}	0.25	**−0.90**	−0.13	
Model M1				
θ_{HCG}				
θ_{HC}	−0.16			
τ_{HCG}	**−0.70**	0.11		
τ_{HC}	−0.08	**−0.82**	0.27	
q	0.06	**0.46**	0.05	**−0.41**

NOTE.—The model of constant mutation rate across loci is used. High correlations are highlighted in bold.

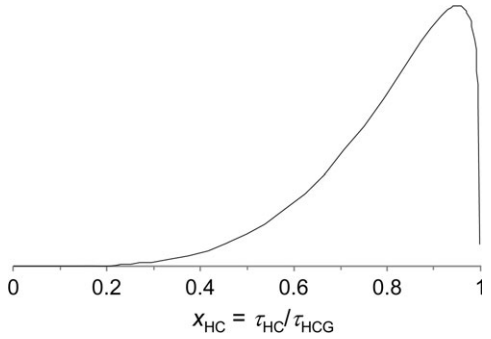

Fig. 3.—The beta distribution for variable τ_{HC} across the genome estimated from the human, chimpanzee, and gorilla genomic sequences (9,681 autosomal loci) under model M1 of variable τ_{HC} among loci.

Discussion

Factors That Cause Variable Species Divergence Times

Here, we discuss several factors that may cause the species divergence time τ to vary over the genome and speculate on their implications to the LRT developed in this paper. First, as discussed in Introduction, gene flow during parapatric or sympatric speciation can cause variation in τ over genomic regions. Similarly, variable τs can be caused by introgression (secondary contact) following allopatric speciation in which reproductive isolation is established without gene flow. It appears very difficult to distinguish between those two scenarios, especially if introgression occurred soon after the initial speciation. No distinction is made between the two in the LRT of this paper. Thus, caution should be exercised in the interpretation of the LRT, as the statistical evidence for variable τs over the genome is compatible with both parapatric speciation with gene flow and allopatric speciation without gene flow followed by secondary contact. In their evaluation of the IM program (Hey and Nielsen 2004), Becquet and Przeworski (2009) considered secondary contact as a version of the null hypothesis of allopatric speciation without gene flow (see their fig. 1*E*) and regarded the detection of gene flow by IM as a false positive error. Here, both parapatric speciation and introgression are considered the alternative hypothesis of speciation with gene flow or different scenarios of the complex speciation model (Patterson et al. 2006).

It is not so clear how τ should vary across the genome when speciation is parapatric and in presence of gene flow. Different models exist that predict the accumulation of genomic incompatibilities over time after one ancestral population splits into two (for reviews, see Turelli et al. 2001; Coyne and Orr 2004; Gourbière and Mallet 2010). Incompatibilities may involve a single locus (i.e., heterozygous disadvantage) or multiple loci. The latter type is known as Dobzhansky–Muller (D-M) incompatibility, which reduces hybrid fitness due to epistatic effects of independent substitutions in different genes since the separation of the two populations. This appears likely to be more important than single-locus incompatibilities. The "snowball" model (Orr 1995; Orr and Turelli 2001) predicts that D-M incompatibilities accumulate at least as fast as τ^2, where τ is the species separation time. The prediction, however, is based on the assumption that many genes are involved in D-M incompatibility and that any pair of genes might interact to create an incompatibility. Different dynamics such as linear accumulation of incompatibilities over time may result from different model assumptions (Kondrashov 2003; Kirkpatrick and Barton 2006; Gourbière and Mallet 2010). In addition to the different predictions of the accumulation of incompatibilities, it is unclear how incompatibilities affect fitness and how the linear or quadratic accumulation of incompatibilities should be translated into a reduction of migration rate and gene exchange over time and to a probability density function $f(\tau)$, which describes the variation of divergence time between the two species across the genome. Intuitively $f(\tau)$ should be single-moded if incompatibilities accumulate gradually, leading to gradual reduction of the migration rate: if t_0 is the inception of species separation and t_1 is the time when gene flow has completely ceased, the density $f(\tau)$ should be >0 in the interval $t_0 < \tau < t_1$ only. It may be noted here that the model of variable τs over loci is only a heuristic approximation to the model of speciation with gene flow, as the process cannot simply be described by variable τs among loci. However, the null model in the LRT is correctly formulated, so that the heuristic nature of the alternative model should affect the power of the LRT but should not cause excessive false positives. A more accurate formulation of the model should consider migrations between the two populations, perhaps with the migration rates changing over time, as well as coalescent events within the two populations and their common ancestor. A Bayesian Markov chain Monte Carlo (MCMC) algorithm appears necessary to implement such a model, by extending the work of Rannala and Yang (2003) and Hey and Nielsen (2004).

In the case of introgression following initial allopatric speciation, as envisaged by Patterson et al. (2006) in their complex speciation model, one should expect τ to have a bimodal distribution. Even though the beta distribution cannot accommodate two modes, it appears appropriate to apply the LRT of this paper to test for introgression against the null hypothesis of a constant τ across the genome.

Another important factor that can cause variable species divergence times over the genome is natural selection. In this regard, it should be noted that the model developed here assumes neutral evolution of gene sequences and may not be suitable for analysis of gene loci under selection. If a locus is under the same purifying selection in different species and the effect is simply to remove strongly deleterious mutations, the strict neutral model may be a reasonable approximation of the evolutionary process at the locus

although with a reduced neutral mutation rate. Most housekeeping genes appear to fit this description as they perform the same function in closely related species and are under similar selective constraints. Use of such genes in the analysis appears justifiable (Ebersberger et al. 2007). The same may apply to neutral loci undergoing background selection because of their linkage to genes under purifying selection (Charlesworth et al. 1993; Nordborg et al. 1996). If the strength of background selection and the recombination rates are similar across species, background selection will have similar effects in different lineages, reducing both diversity and divergence, and the overall effect will be similar to a reduction of mutation rate at the neutral locus.

Although purifying and background selection may have similar effects in different species and thus not cause serious problems to the LRT, positive selection often operates in different ways in different species. For example, ecological adaptations may be highly species specific (Swanson and Vacquier 2002b; Orr et al. 2004). The method developed here is not suitable for analyzing genes under positive selection or genes that cause reproductive isolation or are otherwise involved in the speciation process (Orr et al. 2004; Wu and Ting 2004). Studies of such genes my provide great insights into the speciation process, but their analysis requires different molecular evolutionary tools, such as methods for measuring and testing the strength of positive Darwinian selection (Yang et al. 2000; Swanson and Vacquier 2002a).

Another factor that may cause violations of model assumptions made in the LRT is the population demographic process. Population subdivision in the ancestor may be expected to lead to an increased effective ancestral population size (i.e., large estimates of θ_1) rather than variation in τ and thus may not cause excessive false positives in the LRT. This was the result found by Becquet and Przeworski (2009: fig. 1*C*) in their evaluation of the IM program, and the LRT of this paper may be expected to behave in similar ways. The impact of population size fluctuation such as bottlenecks in the ancestor is less clear: it may likely affect the ancestral population size (θ_1) rather than causing τ to vary among loci.

It may be noted that the conceptual framework of the model of variable τ among loci implemented in this paper is similar to the test of simultaneous species divergences across pairs of sister species, due to a particular geological event, such as the forming of the Isthmus of Panama (Hickerson et al. 2006; Hurt et al. 2009). Such analyses have to overcome similar difficulties such as the confounding effects of variable mutation rates among loci and the strong correlation between the divergence time of the species pair and the ancestral population size. In addition, the ancestral populations of the different sister species have different sizes and separate parameters may have to be used for them. Violation of the molecular clock (i.e., variable rates between the species pairs rather than within each species pair) may complicate the analysis even further. Data of multiple loci from multiple individuals appear necessary to address this problem, although Hickerson et al. (2006) analyzed only one mitochondrial locus and were much more optimistic.

Variable Species Divergence Times and Human–Chimpanzee Speciation

In an analysis of variable sites in the genomes of the human (H), chimpanzee (C), gorilla (G), orangutan (O), and macaque (M), Patterson et al. (2006) suggested that the human–chimpanzee speciation process might have been complex and have involved introgression after the initial separation of the two species. This controversial hypothesis was based on two major pieces of evidence: the large fluctuation of H-C sequence divergence throughout the genome and a dramatic reduction in H-C sequence divergence on the X chromosome. Here, we discuss the implications of the results of this paper to that controversy (see also Barton 2006; Burgess and Yang 2008; Wakeley 2008).

The large fluctuation of H-C divergence could be explained by a large ancestral population size (or large θ_{HC}) (Barton 2006). Indeed, Burgess and Yang (2008) estimated the HC ancestral population to be ~10 times as large as the modern human population, consistent with early estimates (e.g., Takahata and Nei 1985; Hobolth et al. 2007). More generally, θ estimates for ancestral species have been noted to be much larger than for modern species in many species groups (e.g., Satta et al. 2004; Won et al. 2005; Zhou et al. 2007). A number of authors have suggested that population subdivision in the ancestors may have generated the large effective population sizes (e.g., Osada and Wu 2005; Becquet and Przeworski 2007; Zhou et al. 2007). However, there does not appear to be any evidence that most ancestral species were subdivided, whereas modern species are not. Thus, those large estimates of ancestral θs may be a methodological artifact, due to, for example, gene flow around the time of speciation, as suggested by the LRT of this paper for the hominoid data. If the speciation process is often "unclean," the exchange of migrants would cause large variations in the sequence divergence times, leading to large estimates of ancestral θs under models that do not accommodate gene flow.

Yet another explanation is the differential reduction of diversity at neutral loci due to background selection (Charlesworth et al. 1993). McVicker et al. (2009) found that both diversity within the human population and divergence between the human and chimpanzee are reduced at putative neutral sites close to exons and other conserved elements, with greater reduction at sites closer to exons. The authors estimated a 19–26% reduction in human diversity at neutral sites due to background selection. However, background selection may not be very important to the hominoid data analyzed here and by Burgess and Yang (2008) because these data were filtered so that every locus is >1 kb away

from known genes, whereas McVicker et al. (2009) included putatively neutral sites that are often very close to exons. In another study where sites near genes (within 5 kb of transcripts and within 1 kb of exons) were excluded, the estimated reduction in diversity was small (6%) (Cai et al. 2009). Furthermore, the background selection considered by McVicker et al. (2009) should reduce both diversity and divergence, so that its effect should be similar to reduction of the mutation rate for the neutral locus. This effect has been considered and found to be unimportant for the hominoid data by Burgess and Yang (2008).

A second major observation that may be inconsistent with a simple model of human–chimpanzee speciation is the extreme reduction in the H-C sequence divergence (but not in the H-G divergence) on the X chromosome. Note that many factors can contribute to this reduction. 1) The mutation rate is higher in males than in females (Haldane 1935; Li et al. 2002; Ellegren 2007), resulting in a mutation rate difference between the X chromosome and the autosomes ($\mu_x/\mu_A < 1$). 2) The X and A loci have different effective population sizes, with $N_x/N_A = 3/4$ for a 1:1 sex ratio. 3) Processes such as introgression may have caused the H-C species divergence time to differ between the X and autosomal loci. The analysis of Burgess and Yang (2008), under models of constant τ_{HC} across the genome, suggested that the reduction in H-C sequence divergence on the X was mostly due to a reduced population size (θ_{HC}) rather than a reduced species divergence time (τ_{HC}).

To explore the contributions of the various factors to the reduced H-C divergence on the X, we calculated the X/A ratios of θ and τ estimates for the different ancestors (supplementary table S3, Supplementary Material online), following Burgess and Yang (2008). The sensitivity of estimates of parameters such as θ_{HC} to the inclusion or exclusion of orangutan and macaque is intriguing and causes the X/A ratios of θ and τ estimates to depend on the data sets as well. Furthermore, the number of X loci is relatively small, so that parameter estimates for the X chromosome involve considerable sampling errors. One may expect that the HCG and the HCGO data sets are less affected by violations of the molecular-clock assumption or by genomic rearrangements that may alter the neutral mutation rate. For example, the structure of the X chromosome appears to be conserved in all the great apes (Muller and Wienberg 2001; Stanyon et al. 2008). Thus, we focus on the HCG and HCGO data sets and on the large data sets with smaller sampling errors. The τ_x/τ_A ratio was very consistent in the different analyses and data sets, being 0.85 for the HCG ancestor and 0.82 for all others, so the estimate 0.83 used by Burgess and Yang (2008) appears reliable. Note from $\mu_x/\mu_A = 0.82$ and 0.84, one obtains the male/female mutation rate ratio $\alpha = \mu_M/\mu_F = 3.3$ and 2.8, respectively. This consistency implies that the male/female mutation rate most likely stayed constant among the hominoid ancestors (cf. Wakeley 2008). The θ_x/θ_A ratio for the HCG ancestor varies among the data sets used, at about 0.80 in the HCG data sets, 0.70 in the HCGO data sets, and 0.65 in the HCGOM data sets. Divided by the rate ratio $\mu_x/\mu_A = 0.83$, these θ ratios translate to the population size ratios $N_x/N_A = 0.96$, 0.84, 0.78, all higher than the expected 3/4. The θ_x/θ_A ratio for HC is about 0.38–0.40, which implies $N_x/N_A = 0.40$–0.48, much lower than 3/4. Those calculations are affected by the limited number of loci on the X chromosome and the large sampling errors in the θ estimates for the X. Future studies may benefit from including more X loci and from the analysis of the genomic sequences for the Y chromosome from the chimpanzee (Hughes et al. 2010) and other great apes.

Presgraves and Yi (2009) suggest that the variation in male mutation rate among the great apes caused by different mating systems and different intensities of sperm competition may explain the data. Sperm competition is expected to be weak or absent in gorillas and orangutans but intense in chimpanzees with humans to be intermediate. The authors estimated α to be about 2.8–3.6 for humans and 5.1–5.8 for chimpanzees in a data set involving HCGM, consistent with the sperm-competition hypothesis, where estimates of a are 3.3–4.1 for humans and 3.0–3.8 for chimpanzees. Estimates of α for the gorilla and orangutan were around 1.2–1.7 and 1.6–1.8, respectively. However, the authors' estimation procedure is somewhat simplistic. It fixes θs for the different ancestors at the same values and does not account for variation in gene genealogies across the genome. If the hypothesis of sperm competition is true, one would expect the X loci to evolve at more homogeneous rates among lineages, whereas the molecular clock should be violated at the autosomal loci, with the chimpanzee having the highest rate and the gorilla the lowest rate. However, those expectations are not supported by the average sequence distances between those species calculated by Burgess and Yang (2008: table 1). The gorilla had the largest distance (or highest rate) compared with the human and chimpanzee at both the A and X loci, apparently because of the high sequence errors in the gorilla sequence. The human and chimpanzee distances were very close at both the autosomal and X loci ($d_{HO} = 0.0346$ vs. $d_{CO} = 0.0348$ for autosomal loci and $d_{HO} = 0.0278$ vs. $d_{CO} = 0.0277$ for the X loci).

Pool and Nielsen (2007) showed that demographic processes such as population bottlenecks may have disproportional effects on the diversity of autosomal and X-linked loci and that as a result, the N_x/N_A ratio may deviate from the expected 3/4. Thus, a bottleneck in the HC ancestor could cause N_x/N_A to be smaller than 3/4. However, this effect appears small for parameter values reasonable for the hominoids. Furthermore, although bottlenecks are generally acknowledged to have occurred in modern humans when humans migrated out of Africa, there is yet no known evidence for bottlenecks in the HC ancestor, and instead,

the large θ_{HC} estimates for the autosomal loci appear inconsistent with such bottlenecks.

In sum, the process of human–chimpanzee speciation remains poorly understood. The large ancestral population sizes (large θs) may reflect biological reality such as population subdivision in the ancestral species but may also be an artifact of the estimation procedure because of model violations. One important such violation is gene flow around the time of speciation, which elevates the variance in the H-C sequence divergence times and leads to large estimates of ancestral θs. The severely reduced H-C divergence on the X chromosome is intriguing, as is the sensitivity of estimates of certain parameters to the inclusion or exclusion of the orangutan and macaque sequences. Analysis of more data from the X chromosome and of the Y genomic data may shed light on the issue.

Computational Limitation of the Maximum Likelihood Method

The use of maximum likelihood without the need for priors may be considered an advantage of the method. Nevertheless, the current implementation is limited to three species, with one sequence from each. The likelihood computation involves 3-D integrals under model M1, which seems near the limit of computational feasibility. Every additional sequence would mean an extra dimension in the integral. This "curse of dimension" makes it difficult to extend the present model to more species or more sequences. In this regard, Bayesian MCMC methods offer a clear advantage, and it should be straightforward to implement the same model in the framework of Rannala and Yang (2003).

Program availability. A C program (3s) implementing the models of this paper is available at the web site http://abacus.gene.ucl.ac.uk/software/. This replaces the program Ne3sML (Yang 2002).

Supplementary Material

Supplementary tables S1–S3 are available at *Genome Biology and Evolution* online (http://www.oxfordjournals.org/our_journals/gbe/).

Acknowledgments

This paper originated from discussions with Chung-I Wu and benefited from discussions with Jim Mallet. I thank Jim Mallet, David Reich, and two anonymous referees for many constructive comments. Z.Y. is a Royal Society Wolfson Merit Award holder.

Literature Cited

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

Becquet C, Przeworski M. 2007. A new approach to estimate parameters of speciation models with application to apes. Genome Res. 17:1505–1519.

Becquet C, Przeworski M. 2009. Learning about modes of speciation by computational approaches. Evolution. 63:2547–2562.

Burgess R, Yang Z. 2008. Estimation of hominoid ancestral population sizes under Bayesian coalescent models incorporating mutation rate variation and sequencing errors. Mol Biol Evol. 25:1979–1994.

Cai JJ, Macpherson JM, Sella G, Petrov DA. 2009. Pervasive hitchhiking at coding and regulatory sites in humans. PLoS Genet. 5: e1000336.

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.

Coyne JA, Orr HA. 2004. Speciation. Sunderland (MA): Sinauer Associates.

Ebersberger I, et al. 2007. Mapping human genetic ancestry. Mol Biol Evol. 24:2266–2276.

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

Felsenstein J. 2005. Accuracy of coalescent likelihood estimates: do we need more sites, more sequences, or more loci? Mol Biol Evol. 23:691–700.

Gourbière S, Mallet J. 2010. Are species real? The shape of the species boundary with exponential failure, reinforcement, and the "missing snowball". Evolution. 64:1–24.

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

Heled J, Drummond AJ. 2008. Bayesian inference of population size history from multiple loci. BMC Evol Biol. 8:289.

Hey J, Nielsen R. 2004. Multilocus methods for estimating population sizes, migration rates and divergence time, with applications to the divergence of *Drosophila pseudoobscura* and *D. persimilis*. Genetics. 167:747–760.

Hickerson MJ, Stahl EA, Lessios HA. 2006. Test for simultaneous divergence using approximate Bayesian computation. Evolution. 60:2435–2453.

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.

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

Hughes JF, et al. 2010. Chimpanzee and human Y chromosomes are remarkably divergent in structure and gene content. Nature. 463:536–539.

Hurt C, Anker A, Knowlton N. 2009. A multilocus test of simultaneous divergence across the Isthmus of Panama using snapping shrimp in the genus *Alpheus*. Evolution. 63:514–530.

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

Kincaid D, Cheney W. 2002. Numerical analysis: mathematics of scientific computing. Pacific Grove (CA): Brooks/Cole.

Kirkpatrick M, Barton NH. 2006. Chromosome inversions, local adaptation and speciation. Genetics. 173:419–434.

Kondrashov AS. 2003. Accumulation of Dobzhansky-Muller incompatibilities within a spatially structured population. Evolution. 57:151–153.

Li WH, Yi S, Makova K. 2002. Male-driven evolution. Curr Opin Genet Dev. 12:650–656.

Liu L, Yu L, Kubatko L, Pearl DK, Edwards SV. 2009. Coalescent methods for estimating phylogenetic trees. Mol Phylogenet Evol. 53: 320–328.
McVicker G, Gordon D, Davis C, Green P. 2009. Widespread genomic signatures of natural selection in hominid evolution. PLoS Genet. 5:e1000471.
Muller S, Wienberg J. 2001. "Bar-coding" primate chromosomes: molecular cytogenetic screening for the ancestral hominoid karyotype. Hum Genet. 109:85–94.
Nordborg M, Charlesworth B, Charlesworth D. 1996. The effect of recombination on background selection. Genet Res. 67:159–174.
Orr HA. 1995. The population genetics of speciation: the evolution of hybrid incompatibilities. Genetics. 139:1805–1813.
Orr HA, Masly JP, Presgraves DC. 2004. Speciation genes. Curr Opin Genet Dev. 14:675–679.
Orr HA, Turelli M. 2001. The evolution of postzygotic isolation: accumulating Dobzhansky-Muller incompatibilities. Evolution. 55:1085–1094.
Osada N, Wu CI. 2005. Inferring the mode of speciation from genomic data: a study of the great apes. Genetics. 169:259–264.
Patterson N, Richter DJ, Gnerre S, Lander ES, Reich D. 2006. Genetic evidence for complex speciation of humans and chimpanzees. Nature. 441:1103–1108.
Pool JE, Nielsen R. 2007. Population size changes reshape genomic patterns of diversity. Evolution. 61:3001–3006.
Presgraves DC, Yi SV. 2009. Doubts about complex speciation between humans and chimpanzees. Trends Ecol Evol. 24:533–540.
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.
Rannala B, Yang Z. 2008. Phylogenetic inference using whole genomes. Annu Rev Genomics Hum Genet. 9:217–231.
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.
Schwarz G. 1978. Estimating the dimension of a model. Ann Statist. 6:461–464.
Self SG, Liang K-Y. 1987. Asymptotic properties of maximum likelihood estimators and likelihood ratio tests under nonstandard conditions. J Am Stat Assoc. 82:605–610.
Stanyon R, et al. 2008. Primate chromosome evolution: ancestral karyotypes, marker order and neocentromeres. Chromosome Res. 16:17–39.
Swanson WJ, Vacquier VD. 2002a. The rapid evolution of reproductive proteins. Nat Rev Genet. 3:137–144.
Swanson WJ, Vacquier VD. 2002b. Reproductive protein evolution. Annu Rev Ecol Syst. 33:161–179.
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, Nei M. 1985. Gene genealogy and variance of interpopulational nucleotide differences. Genetics. 110:325–344.
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.
Turelli M, Barton N, Coyne J. 2001. Theory and speciation. Trends Ecol Evol. 60:325–413.
Wakeley J. 2008. Complex speciation of humans and chimpanzees. Nature. 452:E3–E4.
Wilson IJ, Weal ME, Balding DJ. 2003. Inference from DNA data: population histories, evolutionary processes and forensic match probabilities. J R Statist Soc A. 166:155–201.
Won YJ, Sivasundar A, Wang Y, Hey J. 2005. On the origin of Lake Malawi cichlid species: a population genetic analysis of divergence. Proc Natl Acad Sci U S A. 102(Suppl 1):6581–6586.
Wu CI, Ting CT. 2004. Genes and speciation. Nat Rev Genet. 5:114–122.
Yang Z. 1994. Statistical properties of the maximum likelihood method of phylogenetic estimation and comparison with distance matrix methods. Syst Biol. 43:329–342.
Yang Z. 1997. PAML: a program package for phylogenetic analysis by maximum likelihood. Comput Appl Biosci. 13:555–556.
Yang Z. 2002. Likelihood and Bayes estimation of ancestral population sizes in Hominoids using data from multiple loci. Genetics. 162: 1811–1823.
Yang Z. 2006. Computational molecular evolution. Oxford: Oxford University Press.
Yang Z, Swanson WJ, Vacquier VD. 2000. Maximum likelihood analysis of molecular adaptation in abalone sperm lysin reveals variable selective pressures among lineages and sites. Mol Biol Evol. 17: 1446–1455.
Zhou R, et al. 2007. Population genetics of speciation in nonmodel organisms: I. ancestral polymorphism in mangroves. Mol Biol Evol. 24:2746–2754.

Associate editor: Hidemi Watanabe

Table S1
Estimates of Parameters under Model M0 from Hominoid Autosomal Loci

Method & Data	L	θ_{HCGO}	θ_{HCG}	θ_{HC}	τ_{HCGO}	τ_{HCG}	τ_{HC}
ML							
HCG, clean	9,861		3.58 ± 0.08	4.31 ± 0.25		6.61 ± 0.04	4.32 ± 0.07
HCG, clean	14,663		3.66 ± 0.07	4.30 ± 0.22		6.62 ± 0.03	4.32 ± 0.06
Bayesian							
HCG, clean [a]	9,861		3.60 ± 0.08	4.19 ± 0.25		6.60 ± 0.04	4.35 ± 0.08
HCG, messy [a]	9,861		3.62 ± 0.09	4.21 ± 0.28		6.61 ± 0.04	4.36 ± 0.08
HCG, clean [a]	14,663		3.67 ± 0.07	4.21 ± 0.21		6.62 ± 0.03	4.34 ± 0.06
HCG, messy [a]	14,663		3.69 ± 0.07	4.23 ± 0.22		6.64 ± 0.03	4.35 ± 0.07
HCGO, clean [b]	14,663	8.05 ± 0.15	3.47 ± 0.06	5.99 ± 0.22	13.68 ± 0.07	6.60 ± 0.03	4.05 ± 0.05
HCGO, messy [b]	14,663	8.08 ± 0.15	3.50 ± 0.06	6.01 ± 0.21	13.75 ± 0.07	6.63 ± 0.03	4.07 ± 0.05
HCGOM, clean [c]	14,663	6.01 ± 0.13	3.36 ± 0.05	6.40 ± 0.21	14.45 ± 0.06	6.68 ± 0.03	4.05 ± 0.05
HCGOM, messy [c]	14,663	6.11 ± 0.14	3.42 ± 0.05	6.40 ± 0.21	14.62 ± 0.07	6.75 ± 0.03	4.11 ± 0.05
Bayesian (BY08)							
HCGOM [d]	14,663	4.9 (4.7-5.2)	3.3 (3.2-3.4)	6.1 (5.7-6.6)	14.3 (14.2-14.5)	6.3 (6.2-6.4)	3.9 (3.8-4.0)

Note .— Estimates of θ and τ are scaled by 10^3. Sites with missing nucleotides or alignment gaps are removed in the "clean" datasets and are included in the "messy" datasets. The ML method is implemented for "clean" data only.

[a] The priors are $\theta \sim G(2, 2000)$ with mean 0.001, and $\tau_{HCG} \sim G(2, 300)$ with mean 0.0067.

[b] The priors are $\theta \sim G(2, 2000)$ with mean 0.001, and $\tau_{HCGO} \sim G(2, 120)$ with mean 0.0167.

[c] The priors are $\theta \sim G(2, 2000)$ with mean 0.001, and $\tau_{HCGOM} \sim G(2, 80)$ with mean 0.025.

[d] The posterior means and 95% CIs from table 2 "(d) random-rates model" of Burgess and Yang (2008), obtained using MCMCcoal1.2. These are quoted here as they are the best estimates from that study. The results from the basic model (Burgess and Yang 2008: table 2a) are virtually identical to the Bayesian estimates from the BCGOM messy data. The posterior distribution is nearly normal and the SD is roughly ¼ times the 95% posterior CI width.

The new Bayesian analyses is conducted using different and mostly more diffuse priors than in Burgess and Yang (2008). The τ for the root of the tree is assigned a gamma prior while other τs are assigned a uniform Dirichlet prior given the root τ.

Table S2
Estimates of Parameters under Model M0 from Hominoid X-Chromosome Loci

Method & Data	L	θ_{HCGO}	θ_{HCG}	θ_{HC}	τ_{HCGO}	τ_{HCG}	τ_{HC}
ML							
HCG, clean	510		3.05 ± 0.38	1.42 ± 0.47		5.21 ± 0.19	3.62 ± 0.23
HCG, clean	783		3.00 ± 0.31	1.62 ± 0.37		5.40 ± 0.16	3.54 ± 0.18
Bayesian							
HCG, clean	510		2.87 ± 0.37	1.40 ± 0.41		5.28 ± 0.19	3.66 ± 0.21
HCG, messy	510		2.88 ± 0.38	1.43 ± 0.43		5.29 ± 0.20	3.65 ± 0.22
HCG, clean	783		2.90 ± 0.31	1.58 ± 0.35		5.43 ± 0.16	3.57 ± 0.17
HCG, messy	783		2.89 ± 0.30	1.63 ± 0.35		5.46 ± 0.16	3.55 ± 0.17
HCGO, clean	783	5.23 ± 0.53	2.45 ± 0.24	2.32 ± 0.33	11.50 ± 0.27	5.60 ± 0.14	3.33 ± 0.14
HCGO, messy	783	5.22 ± 0.54	2.44 ± 0.25	2.38 ± 0.33	11.56 ± 0.28	5.64 ± 0.14	3.34 ± 0.14
HCGOM, clean	783	3.64 ± 0.54	2.18 ± 0.22	2.53 ± 0.31	12.09 ± 0.29	5.75 ± 0.13	3.33 ± 0.14
HCGOM, messy	783	3.58 ± 0.53	2.22 ± 0.23	2.57 ± 0.31	12.32 ± 0.29	5.84 ± 0.13	3.37 ± 0.13
Bayesian (BY08)							
HCGOM	783	3.2 (1.9-4.4)	2.0 (1.6-2.5)	2.6 (2.0-3.3)	11.7 (11.1-12.3)	5.4 (5.1-5.6)	3.1 (2.8-3.9)

Note.— The same priors are used as in table S1. See legend to table S1.

Table S3
The X/A Ratios of θs and τs for Ancestors HC and HCG

Method & Data	L	θ_{HCGO}	θ_{HCG}	θ_{HC}	τ_{HCGO}	τ_{HCG}	τ_{HC}
ML							
HCG, clean	small		0.852	0.329		0.788	0.838
HCG, clean	large		0.820	0.377		0.816	0.819
Bayesian							
HCG, clean	small		0.797	0.334		0.800	0.841
HCG, messy	small		0.796	0.340		0.800	0.837
HCG, clean	large		0.790	0.375		0.820	0.823
HCG, messy	large		0.783	0.385		0.822	0.816
HCGO, clean	large	0.650	0.706	0.387	0.841	0.848	0.822
HCGO, messy	large	0.646	0.697	0.396	0.841	0.851	0.821
HCGOM, clean	large	0.606	0.649	0.395	0.837	0.861	0.822
HCGOM, messy	large	0.585	0.649	0.402	0.842	0.865	0.820
HCGOM (BY08)	large	0.653	0.606	0.426	0.818	0.857	0.795

Note.— The point estimates of tables S1 and S2 are used to calculate the ratios. The small datasets include 9,861 autosomal loci and 510 X loci, with loci for which the orangutan is missing removed. The large datasets include 14,663 autosomal loci and 783 X loci.