An Autosomal Analysis Gives No Genetic Evidence for Complex Speciation of Humans and Chimpanzees

Masato Yamamichi,¹ Jun Gojobori,¹ and Hideki Innan^{*1,2}

¹Department of Evolutionary Studies of Biosystems, Graduate University for Advanced Studies, Hayama, Kanagawa, Japan

²PRESTO, Japan Science and Technology Agency (JST), Saitama, Japan

***Corresponding author:** E-mail: innan.hideki@soken.ac.jp.

Associate editor: Rasmus Nielsen

Abstract

There have been conflicting arguments as to what happened in the human–chimpanzee speciation event. Patterson et al. (2006, Genetic evidence for complex speciation of humans and chimpanzees. *Nature* 441:1103–1108) proposed a hypothesis that the human–chimpanzee speciation event involved a complicated demographic process: that is, the ancestral lineages of humans and chimpanzees experienced temporal isolation followed by a hybridization event. This hypothesis stemmed from two major observations: a wide range of human–chimpanzee nucleotide divergence across the autosomal genome and very low divergence in the X chromosome. In contrast, Innan and Watanabe (2006, The effect of gene flow on the coalescent time in the human–chimpanzee ancestral population. *Mol Biol Evol.* 23:1040–1047) demonstrated that the null model of instantaneous speciation fits the genome-wide divergence data for the two species better than alternative models involving partial isolation and migration. To reconcile these two conflicting reports, we first reexamined the analysis of autosomal data by Patterson et al. (2006). By providing a theoretical framework for their analysis, we demonstrated that their observation is what is theoretically expected under the null model of instantaneous speciation with a large ancestral population. Our analysis indicated that the observed wide range of autosomal divergence is simply due to the coalescent process in the large ancestral population of the two species. To further verify this, we developed a maximum likelihood function to detect evidence of hybridization in genome-wide divergence data. Again, the null model with no hybridization best fits the data. We conclude that the simplest speciation model with instantaneous split adequately describes the human–chimpanzee speciation event, and there is no strong reason to involve complicated factors in explaining the autosomal data.

Key words: speciation, isolation, hybridization, coalescent, human, chimpanzee.

Introduction

Speciation is a major topic in evolutionary biology (Coyne and Orr 2004). With the recent availability of enormous amounts of genome sequence data, it is now possible to use a new approach to study speciation: that is, comparative genomics. One of the most interesting models would be the speciation of humans and chimpanzees, our closest living relatives. The human–chimpanzee speciation event is undoubtedly of special interest to us (e.g., Siepel 2009). Solving this problem would answer an important question: what makes us humans?

Two studies of human–chimpanzee speciation, published at almost the same time, have seemingly conflicting conclusions. Innan and Watanabe (2006) showed that the divergence between the two species is very well explained under the simplest speciation model, which assumes that the ancestral population of the two species split instantaneously at a single time point. In contrast, Patterson et al. (2006) suggested that human–chimpanzee speciation involved complex events, including isolation followed by hybridization.

Our aim is to reconcile the findings of these two studies. The major question is why such different conclusions were derived for the same event when part of the analyzed data was shared by the two studies. To answer this

question, we carefully review and reexamine the analysis by Patterson et al. (2006) because several drawbacks in their analysis have been pointed out (Barton 2006; Takahasi and Innan 2008; Wakeley 2008; Presgraves and Yi 2009). Patterson et al. (2006) proposed their hybridization hypothesis on the basis of two major observations: i) there is substantial variation in the divergence between human and chimpanzee across the autosomes, and a range of ~4 million years (My) in the time to the human–chimpanzee common ancestor is required to explain the observed variation and ii) the divergence of the X chromosome is very low in comparison with that of the autosomes, and the variation in divergence for the X chromosome is much less than that for the autosomes. On the basis of these observations, Patterson et al. (2006) suggested that the human–chimpanzee speciation process involved temporal isolation followed by a period of admixture (migration and hybridization): that is, an initial divergence of the two lineages occurred first, and then they interbred again and exchanged genetic information before the two species became completely separated. Patterson et al. (2006) proposed that the observed small amount of variation in divergence across the X chromosome can be explained if the lineages leading to humans and chimpanzees in regard to the X chromosome met only in the admixture period.

The complex human–chimpanzee speciation hypothesis proposed by Patterson et al. (2006) is highly controversial. With regard to the autosomes, Barton (2006) pointed out that the argument of Patterson and colleagues lacks statistics to show that the observed wide range of time to the human–chimpanzee common ancestor is sufficiently large to reject the simplest speciation model, which assumes an instantaneous split of the ancestral population. The effective size of the human–chimpanzee ancestral population is estimated to have been quite large, probably 50,000–100,000 (Takahata et al. 1995; Takahata and Satta 1997; Chen and Li 2001). If this was so, then the observed wide range of divergence can be accounted for in the null model of instantaneous speciation (Takahasi and Innan 2008). Although Patterson and colleagues' nonparametric analysis is straightforward and intuitively appealing, there is no theoretical basis behind their analysis, and it is difficult to understand what it really means.

Patterson and colleagues' observation in regard to the X chromosome is strange and puzzling, but it has been pointed out that the hybridization scenario may not be the only explanation. Barton (2006) provided several alternative scenarios involving differences in mode of selection and mutation rate between the autosomes and the X chromosome (see also Wakeley 2008; Presgraves and Yi 2009). Takahasi and Innan (2008) questioned whether the data used by Patterson et al. (2006) were in fact representative of the whole X chromosome because the data cover only 0.24% of this chromosome.

Here, we carefully reexamine the analysis by Patterson et al. (2006), focusing particularly on the analysis of autosomal data because the autosomes are where we would most likely find footprints of any historical event that happened to the species. This is especially true when genome-wide sequence data are available. Innan and Watanabe (2006) have already shown that the coalescent pattern of the autosomes is best explained by a very simple model of an instantaneous split rather than an alternative model that allows a period of gene flow. To confirm that this conclusion also holds with another alternative model (i.e., hybridization), we reanalyze the autosomal data by modifying the maximum likelihood (ML) method developed by Innan and Watanabe (2006). It is widely accepted that the evolution of sex chromosomes is much more complicated than that of autosomes, and that it involves a number of less understood evolutionary problems, including male–female mutation rate differences and the effect of selection on hemizygotes. Therefore, it is difficult to argue about the speciation event from sex chromosome data, but we attempt to discuss potential problems in Patterson and colleagues' analysis of the X chromosome.

Autosome

We first review Patterson and colleagues' analysis of the autosomes from which they claimed a wide range of divergence time across the autosomes. Although their nonparametric analysis is intuitive, it is difficult to understand what their analysis means due to a lack of theory behind

their analysis. Here, we provide a coalescent-based theory to understand their analysis and show that the result of their analysis is exactly what we expect under the simplest speciation model with an instantaneous split. Furthermore, we develop an ML function to detect hybridization in the ancestral population. Its application to the human–chimpanzee genomic data shows that there is no evidence for hybridization.

Autosomal Analysis by Patterson et al. (2006)

Patterson and colleagues' first observation was that the divergence between humans and chimpanzees varies greatly across the autosomes. On the basis of their nonparametric analysis of autosomes, they estimated that the range of divergence is >4 My (see fig. 2 in Patterson et al. 2006). In this figure, the human–chimpanzee divergence (τ) is compared between regions near an HC site and those near an HG or CG site. H, C, and G represent human, chimpanzee, and gorilla, respectively. An HC site is a site at which the configuration of nucleotides in the alignment of the human, chimpanzee, and gorilla genomes is 110, where 0 and 1 represent the ancestral and derived allelic states, respectively; the patterns at HG and CG sites are 101 and 011. Following Patterson et al. (2006), let τ_{genome} be the genome average of the divergence time, and let A represent $\tau/\tau_{\text{genome}}$, the relative divergence time in a local region. Patterson et al. (2006) found that τ near HC sites is $<90\%$ of the genomic average (i.e., $A \approx 0.9$), whereas τ near HG or CG sites is about 45% larger than τ_{genome} (i.e., $A \approx 1.45$). However, such a wide range of divergence is not a new observation. Classically, Takahata and colleagues (Takahata et al. 1995; Takahata and Satta 1997) reported a similar finding and suggested that such variation could be explained by the stochastic variation in coalescent time in a large ancestral population. They estimated the effective size of the ancestral population to be 50,000–100,000, which is several times larger than that of the current human population (Takahata et al. 1995; Takahata and Satta 1997).

More importantly, we here show that Patterson and colleagues' observation is exactly what we expect under the null model of instantaneous speciation. Because Patterson et al. (2006) focused on sites at which there is variation among humans, chimpanzees, and gorillas, the theory required to understand their analysis is the coalescent theory at a biallelic site (see Innan and Tajima 1997; Wiuf and Donnelly 1999 for theory). We first briefly describe this theory in a simple situation in which the ancestral relationships of three lineages are considered in a panmictic population and then we extend it to the human–chimpanzee–gorilla case.

Consider a single nucleotide site at which we know that there are two alleles. It is assumed that this biallelic polymorphism originates from a single mutational event. This should be a reasonable assumption in most cases of nucleotide polymorphism because the mutation rate per site is very low (Kimura 1969). Figure 1A illustrates a simple situation in a single diploid population of constant size with sample size $n = 3$, represented by X, Y, and Z. Consider the coalescent process of X, Y, and Z backward in time. The first

FIG. 1. (A) Standard coalescent process with three haploid individuals, X, Y and Z. (B) Coalescent process at a biallelic site. The black circle represents the mutation that created two alleles. (C) Simple allopatric speciation model for human, chimpanzee, and gorilla. Typical genealogical trees at HC and CG sites are also illustrated with blue and red lines. (D) The effect of recombination on the relative age, A . The gray lines show the simulation results in the null model of instantaneous speciation. The observation of Patterson et al. (2006) is shown by filled squares (data from figure 2 in Patterson et al. 2006).

event is the coalescence of two of the three lineages, say Y and Z, at time t_2 . The standard coalescent theory (Kingman 1982; Hudson 1983; Tajima 1983) predicts that the expectation of t_2 is $2N/3$ generations, where N is the effective population size. Then, the ancestor of Y and Z coalesces with X at time $t_1 + t_2$, and the expectation of t_1 is $2N$ generations. However, this standard theory does not hold when we know that the site has two alleles, 0 and 1. Figure 1B illustrates such a case, where X and Y have the derived state 1 and Z has the ancestral state 0. The prior knowledge restricts the coalescent process: First, X and Y must coalesce, and then the common ancestor of X and Y coalesces with Z. The expectation of the coalescent time (t_2) of X and Y is still $2N/3$, but the expectation of t_1 under this condition is $4N$ —twice that under the standard coalescent—because a mutation has to be involved during this coalescent time. This is intuitively difficult to understand but mathematically true. For the X–Y common ancestor and Z to coalesce, they must first wait for a mutation to occur. The waiting time for the mutation follows a simple function of the mutation rate, and Tajima (1983) proved that this becomes an exponential

distribution with mean $2N$ generations when the mutation rate per generation per site is very low. Once this mutation occurs, the two lineages have no restriction on coalescence, so the coalescent time follows an exponential distribution with mean $2N$ generations. Thus, the entire process of the coalescence of the X–Y common ancestor and Z involves two independent processes, each of which follows the same exponential function. Hence, the probability distribution of the coalescent time conditional on the presence of a mutation is given by the convolution of two exponential distributions and the mean is given by $4N$. Recently, Huff et al. (2010) developed a unique algorithm to estimate ancestral population size by taking advantage of this somewhat complex theoretical property.

It is straightforward to extend this theory to the instantaneous speciation model of humans, chimpanzees, and gorillas (fig. 1C). The model requires four parameters, two speciation times, $\tau_{\text{species}}^{\text{HC}}$ and $\tau_{\text{species}}^{\text{HG}}$, where $\tau_{\text{species}}^{\text{HG}} = \tau_{\text{species}}^{\text{HC}} + T_1$, and two ancestral population sizes, N_{HC} and N_{HCG} , where $N_{\text{HCG}} = kN_{\text{HC}}$. For convenience, time is measured in units of $2N_{\text{HC}}$ generations. Note that for this theory,

the population sizes after speciation do not matter because there is no recombination (see below for a case with recombination). **Figure 1C** is a typical genealogical tree at an HC site (blue) and that at a CG site (red). The relationship between human and chimpanzee in the former case corresponds to that of X and Y in **figure 1B**, whereas the situation in the latter case is similar to that of X and Z, or Y and Z, in the same figure. Therefore, it is obvious that the time to the most recent common ancestor (TMRCA) of human and chimpanzee should be, on average, less at an HC site (blue tree) than that at an HG or CG site (red tree). This is because the coalescence of the human and chimpanzee lineages involves a mutation at an HG or CG site but not at an HC site.

First, the expected TMRCA of human and chimpanzee is considered under the standard coalescent, which is denoted by τ_{genome} . It is quite straightforward to obtain τ_{genome}

$$\begin{aligned}\tau_{\text{genome}} &= \tau_{\text{species}}^{\text{HC}} + \int_0^{T_1} t e^{-t} dt + (T_1 + k) \int_{T_1}^{\infty} e^{-t} dt \\ &= 1 + \tau_{\text{species}}^{\text{HC}} + (k - 1)e^{-T_1}.\end{aligned}\quad (1)$$

The term $\int_0^{T_1} t e^{-t} dt$ accounts for the coalescent process in the human–chimpanzee ancestral population, and the term $(T_1 + k) \int_{T_1}^{\infty} e^{-t} dt$ accounts for the process in the ancestral population of all three species. Note that equation (1) can be applied to any site with no prior information.

Information on the presence of a mutation can be incorporated as described with **figure 1B**. τ_{genome} at an HC site is given by the following equation:

$$\begin{aligned}\tau_{\text{genome}}^{\text{HC}} &= \tau_{\text{species}}^{\text{HC}} + \int_0^{T_1} t e^{-t} dt + (T_1 + k/3) \int_{T_1}^{\infty} e^{-t} dt \\ &= 1 + \tau_{\text{species}}^{\text{HC}} + \frac{k - 3}{3} e^{-T_1}.\end{aligned}\quad (2)$$

Next, consider the genealogical tree at an HG or CG site. The human and chimpanzee lineages cannot coalesce between $\tau_{\text{species}}^{\text{HC}}$ and $\tau_{\text{species}}^{\text{HG}}$ (**fig. 1C**), so $\tau_{\text{genome}}^{\text{HG}}$ and $\tau_{\text{genome}}^{\text{CG}}$ can be simply given by the following equation:

$$\tau_{\text{genome}}^{\text{HG}} = \tau_{\text{genome}}^{\text{CG}} = 1 + \tau_{\text{species}}^{\text{HC}} + T_1 + \frac{7k}{3}.\quad (3)$$

These three equations prove that $\tau_{\text{genome}}^{\text{HG}} > \tau_{\text{genome}} > \tau_{\text{genome}}^{\text{HC}}$ for any positive value of k . From equations (1)–(3), it is possible to calculate the ratios of the divergence times at HC and HG sites to the genomic average: $A^{\text{HC}} = \tau_{\text{genome}}^{\text{HC}} / \tau_{\text{genome}}$ and $A^{\text{HG}} = \tau_{\text{genome}}^{\text{HG}} / \tau_{\text{genome}}$, given N_{HC} , k , T_1 , and $\tau_{\text{species}}^{\text{HC}}$. To be roughly consistent with recent analysis (Yu et al. 2004; Becquet et al. 2007), we assume $N_{\text{HCG}} = N_{\text{HC}} = 4.5N_{\text{H}}$, $\tau_{\text{species}}^{\text{HC}} = 24N_{\text{H}}$, and $\tau_{\text{species}}^{\text{HG}} = 28N_{\text{H}}$. This setting translates $k = 1$, $\tau_{\text{genome}}^{\text{HC}} = 2.97$, and $T_1 = 0.22$ in units of $2N_{\text{HC}}$. From these values, we calculated that $A^{\text{HC}} = 0.85$ and $A^{\text{HG}} = 1.42$, in very good agreement with the analysis of Patterson et al. (2006) (**fig. 1D**).

It should be noted that the above theory can be applied to only a very narrow region around the focal biallelic site so that recombination can be ignored. Recombination breaks down the genealogical relationship more the farther one moves away from the biallelic site. This effect of recombination is demonstrated in **figure 2** of Patterson et al. (2006), which shows the average divergence (A) as a function of the distance from the focal biallelic site. The A value approaches 1 as the distance increases. This observation can be reproduced by a simple coalescent simulation of the null model of instantaneous speciation, by using the model shown in **figure 1C** with the parameters used for the theoretical analysis above. This simulation requires N_{H} , N_{C} , and N_{G} to be the current effective population sizes of the three species. We assume $N_{\text{C}} = N_{\text{G}} = 2.5N_{\text{H}}$ on the basis of recent polymorphism data (Yu et al. 2004; Becquet et al. 2007). We set the population mutation and recombination parameters to be $4N_{\text{H}}\mu = 4N_{\text{H}}r = 0.00075$ per site to be roughly consistent with estimates from human populations (Bouffard et al. 1997; International Human Genome Sequencing Consortium 2001), where μ and r are the mutation and recombination rates per generation per site. To simplify the situation, the simulation assumes that multiple mutations at a single site are not allowed, and that the ancestral states of divergent sites are known. Therefore, the simulation result should be compared with the results corrected for multiple mutations reported in Patterson et al. (2006). DNA variation among the three species in a 100-kb region was then simulated for 10,000 times, and the same analysis as that performed by Patterson et al. (2006) was applied (**fig. 1D**). The results of the simulation (gray lines in **fig. 1D**) were very similar to the observations made by Patterson and colleagues (filled squares in **fig. 1D**). For a small distance from an HC site, the average τ was about 15% lower than τ_{genome} , whereas the average τ around an HG or CG site was about 40% larger than τ_{genome} . Thus, the results of the simulation agreed with the results based on the coalescent theory at a biallelic site in which $0.85\tau_{\text{genome}}$ is predicted at an HC site and $1.42\tau_{\text{genome}}$ is predicted at an HG or CG site (see previous section). The deviation from τ_{genome} decreased with increasing distance from the biallelic site owing to recombination. We propose that this trend is not specific to the parameter set used here because it is expected from the theory (eqs. 2–3). Thus, the results shown in **figure 2** of Patterson et al. (2006) are exactly those expected on the basis of the null model of instantaneous speciation.

ML Framework to Detect Hybridization

Next, we examined one of the very important predictions by Patterson et al. (2006), who stated “it (hybridization) might result in a bimodal distribution of τ .” We developed an ML function to detect hybridization by modifying our previous model (Innan and Watanabe 2006), which was designed to detect ancient gene flow, and we applied the function to human–chimpanzee divergence data. To test the prediction using a larger data set than that analyzed by Patterson et al. (2006), the ML

FIG. 2. (A) Model of isolation and hybridization used in this study. (B–C) Distributions of coalescent time predicted under this model (computed by eq. (4)) with various parameters.

function was developed to take full advantage of the availability of the entire genome sequences of human and chimpanzee.

The general two-species model considered in this section is illustrated in figure 2A. It is assumed that the ancestral diploid population is initially panmictic and that the effective population size is constant, N . Time (t) is measured in units of $2N$ generations from present ($t = 0$). Between $t = \tau_{isolation}$ and $t = \tau_{hybrid}$, the ancestral population is split into two subpopulations with sizes iN and $(1-i)N$ ($0 < i < 1$). Then, the population becomes panmictic again until it splits into two completely isolated species, I and II, at $t = \tau_{species}$. The durations of the isolation and hybridization periods are denoted as $T_I = \tau_{isolation} - \tau_{hybrid}$ and $T_H = \tau_{hybrid} - \tau_{species}$, respectively, and T_S is the time since the speciation event. The population sizes of the two subpopulations, I and II, are assumed to be iN and $(1-i)N$, respectively. It should be noted that i is a parameter that determines the ratio of the two subpopulations before T_S , and there is no assumption of the population sizes for the period $0 \leq t \leq T_S$ in which no coalescent event occurs, provided that we consider only a pair of lineages from the two species.

Suppose that we are interested in the evolutionary history of two sequences, one from species I and the other from species II. The number of nucleotide differences between them is determined by the mutation rate and the TMRCA. The latter consists of the time after complete speciation (T_S) and the coalescent time in the ancestral population. Assuming neutrality, it is straightforward to derive the probability density function of the coalescent time, $P(t) = P(t|N, T_S, T_H, T_I, i)$:

$$P(t) = \begin{cases} e^{T_S - t} & \text{for } \tau_{species} < t < \tau_{hybrid}, \\ e^{-T_H} \left\{ i e^{\frac{T_S + T_H - t}{i}} + (1-i) e^{\frac{T_S + T_H - t}{1-i}} \right\} & \text{for } \tau_{hybrid} < t < \tau_{isolation}, \\ e^{T_S + T_I - t} \left\{ 2i(1-i) + i^2 e^{\frac{-T_I}{i}} + (1-i)^2 e^{\frac{-T_I}{1-i}} \right\} & \text{for } \tau_{isolation} < t. \end{cases} \quad (4)$$

The top equation in (4) relates to the coalescent event that occurs in the time interval $\tau_{species} < t < \tau_{hybrid}$ in which there is no restriction to the coalescent event, so it follows an exponential distribution with mean 1 in units of $2N$ generations. If the two lineages do not coalesce during this time interval, which occurs with probability e^{-T_H} , the system enters the next phase $\tau_{hybrid} < t < \tau_{isolation}$ in which there are two isolated subpopulations. In this phase, coalescence can occur only when the two ancestral lineages are in the same subpopulation, which is described in the middle equation in (4). If coalescence still does not occur, which happens with probability $e^{-T_H} \left\{ 2i(1-i) + i^2 e^{\frac{-T_I}{i}} + (1-i)^2 e^{\frac{-T_I}{1-i}} \right\}$, the process enters the last phase $t > \tau_{isolation}$, where the time to the coalescent event follows an exponential distribution with mean 1, as described in the bottom equation in (4).

Patterson et al. (2006) predicted a bimodal distribution of t with a reasonably long term of isolation, say in the order of a million years. Equations (4) indicate that their prediction is true. The distribution has two peaks, the first at $t = \tau_{species}$ and the second at $t = \tau_{isolation}$. Figures 2B and C demonstrate the effects of $\tau_{isolation}$ and τ_{hybrid} on the distribution of coalescent time, assuming $\tau_{species} = 2N$ generations. The

gray line denotes the case of no hybridization and follows an exponential distribution for $t > 1$. When τ_{hybrid} is assumed to be very close to τ_{species} (i.e., $T_H = 0.01$), the distribution has a clear second peak at $\tau_{\text{isolation}}$ (fig. 2B). A very small T_H means a very short period of gene flow, and the situation is close to the hybridization event due to a single episode of gene flow, suggested by Patterson et al. (2006). In figure 2C, the duration of the isolation is fixed at $2N$ generations (i.e., $T_I = 1$). As the duration of hybridization, T_H , increases, the height of the second peak decreases (fig. 2C).

We then examined whether, as suggested by Patterson et al. (2006), a bimodal distribution of the coalescent times of human and chimpanzee fits the data better than a simple exponential distribution predicted under the null model with no hybridization. For this purpose, we developed an ML function of the observed divergence, d , based on the divergence time given by equation (4):

$$P(d|\theta, T_S, T_H, T_I, i) = \int_{T_S}^{\infty} \frac{(2\mu tL)^d \exp(-2\mu tL)}{d!} P(t|N, T_S, T_H, T_I, i) dt. \quad (5)$$

where d represents the number of mutations accumulated during the coalescent process in a region with length L (bp) and follows a Poisson distribution; μ is the mutation rate per generation per site; and θ is the population size-scaled mutation rate, which is given by $\theta = 4N\mu$.

Application to Human–Chimpanzee Divergence Data

The ML function (eq. 5) was applied to human–chimpanzee divergence data, which were downloaded from the Genome Browser database of the University of California, Santa Cruz (UCSC, 2003–2011, <http://hgdownload.cse.ucsc.edu/downloads.html>). We used the pairwise alignment data of the human (hg18, Mar. 2006, NCBI Build 36.1) and chimpanzee (panTro2, Mar. 2006, Washington University Build 2 Version 1) genomes. These data consist of alignments of homologous regions, including orthologous and nonorthologous (most likely paralogous) regions. Because the coalescent theory can be applied only to orthologous divergence, we screened out nonorthologous alignments according to the synteny information in the UCSC database. Then, the remaining putatively orthologous alignment data were further edited by using the procedure of Patterson et al. (2006). From the orthologous alignments, we excluded the following:

- segmental duplications, according to the data of recent analysis (Cheng et al. 2005, Segmental Duplication Database), downloaded from <http://humanparalogy.gs.washington.edu/build36/build36.htm> and <http://humanparalogy.gs.washington.edu/panTro2wgac/panTro2wgac.html>;
- repetitive sequences, identified by using Tandem Repeat Finder (Tandem Repeat Finder, 1999–), downloaded from <http://tandem.bu.edu/trf/trf.html> (Benson, 1999); and
- CpG dinucleotides because of their high mutation rates.

After these screening processes, we obtained putatively orthologous alignments of autosomal regions with a total length of 2.02 Gbp. The nucleotide divergence in these data is 0.0089, which is considerably lower than the well-known divergence value (0.0123) because of the exclusion of the highly mutable CpG dinucleotides (Chimpanzee Sequencing and Analysis Consortium 2005).

To apply the ML function (eq. 5) to this data set, we randomly extracted a huge number of small regions, each of which included 100 bp of aligned non-CpG nucleotides. We used 100-bp regions because equation (5) can be applied only to independent regions within which there is no recombination, and according to our previous experience (Innan and Watanabe 2006), data of 100-bp regions would have reasonable power to resolve the historical events in the two species' ancestral population, while minimizing the effect of recombination. This is the reason why we used 100-bp regions, which could be determined by the tradeoff of the amount of information and the effect of recombination. We later verified this assumption by simulations (see next section). To avoid correlation between neighboring regions, we included only regions that were separated from a neighboring region by at least 5 kb; therefore, we can consider that the regions are almost independent of each other. The ML function (5) was applied to $\sim 330,000$ random 100-bp regions obtained as described above.

Then, we calculated the distribution of the number of nucleotide differences (d), which is denoted by $\mathcal{D}_A = \{\delta_0, \delta_1, \delta_2, \dots, \delta_{10}\}$, where δ_j is the number of fragments with $d = j$. We removed very few fragments with a number of nucleotide differences (d) more than 10 because they might not have been orthologs. The log likelihood of $\mathcal{D}_A$ was computed by using the following equation:

$$\ln L(\theta, T_S, T_H, T_I, i | \mathcal{D}_A) = \sum_{d=0}^{10} \delta_d \ln P(d|\theta, T_S, T_H, T_I, i, d \leq 10), \quad (6)$$

where

$$P(d|\theta, T_S, T_H, T_I, i, d \leq 10) = P(d|\theta, T_S, T_H, T_I, i) / \sum_{j=0}^{10} P(j|\theta, T_S, T_H, T_I, i). \quad (7)$$

The value of $\ln L(\theta, T_S, T_H, T_I, i | \mathcal{D}_A)$ for various parameter settings was calculated, and the results are summarized in figure 3A. First, we fixed $i = 0.2$ and let the other four parameters vary freely. Our somewhat arbitrary choice of $i = 0.2$ roughly reflects the ratio of the current effective population sizes of humans and chimpanzees (Innan and Watanabe 2006) because the effective population size of chimpanzees may be several times larger than that of humans (we will later check the effect of i by using other values of i ; Kaessmann et al. 1999; Kaessmann et al. 2001; Kitano et al. 2003). θ and T_S were changed with increments of $\times 10^{-5}$ and 10^{-2} , respectively. For the durations of the periods of isolation and hybridization, we set $T_I = \{0, 0.01, 0.1, 0.25, 0.5, 1, 2\}$ and

FIG. 3. (A) ML analysis of the autosomal data, $\mathcal{D}_A$, computed by equation (6). The log-likelihood value is scaled so that the maximum log likelihood over all investigated parameter ranges is 0. (B) The observed distribution of d (i.e., $\mathcal{D}_A$), compared with its expectation under the null model with the best-fit parameter.

$T_H = \{0.01, 0.1, 0.2, 0.3, \dots, 1.0\}$. For each pair of T_H and T_I values, we obtained the profile likelihood, that is, the ML conditional on T_H and T_I ($\max[\ln L | T_H, T_I]$). We found that the ML over all parameter ranges ($\max[\ln L] = -418, 232$) was obtained when there was no hybridization (i.e., $T_H = 0$, $\theta = 0.00405$, $T_S = 1.20$), indicating that the null model with instantaneous split best explains the autosomal data, $\mathcal{D}_A$. Figure 3A shows the relative profile likelihood, $\max[\ln L | T_H, T_I] - \max[\ln L]$. For all values of T_I , the highest likelihood is given when $T_H = 0$. The likelihood immediately drops at $T_H = 0.01$ and then gradually increases again with increasing T_H . It is predicted that $\max[\ln L | T_H, T_I]$ approaches $\max[\ln L]$ when T_H is further increased. This pattern is consistent with the distribution of coalescent time; a very small T_H creates a clear bimodal distribution when T_I is reasonably large. The fit becomes worse as T_I increases, indicating that the data cannot be explained well by the bimodal distribution of coalescent time predicted by the hybridization model. This result is consistent with that of Innan and Watanabe (2006) but not with that of Patterson et al. (2006). We compared the observed distribution of d and the distribution expected under the null model with $\theta = 0.00405$ and $T_S = 1.20$ (fig. 3B). The two distributions are very similar, suggesting that the null model explains the data very well. Almost identical results were obtained for $i = \{0.05, 0.1, 0.5\}$ (data not shown). Our estimate of $T_S = 1.20$ corresponds to 6.1 My if the generation time is 20 years and the mutation rate is 8.0×10^{-9} at non-CpG sites. This mutation rate was assumed because we calculated the average divergence to be 0.0089, which

is about 72% of the divergence including CpG sites (0.0123) (Chimpanzee Sequencing and Analysis Consortium 2005). If the mutation rate per generation is 1.1×10^{-8} (Roach et al. 2010), then the rate at non-CpG sites would be 8.0×10^{-9} .

It should be noted that this general hybridization model is not exactly the same as the model suggested by Patterson et al. (2006) in which unidirectional gene flow from the chimpanzee lineage to the human lineage is considered. To check how this difference affects our conclusion, we repeated the same analysis using a model that is more consistent with that of Patterson et al. (2006). We obtained a very similar result with no strong evidence for hybridization (See Supplementary Material online for details).

Power and Robustness of the ML Approach

The power of our ML approach and the robustness of the approach to factors ignored in the ML function were explored by computer simulation. The power to detect hybridization was evaluated by calculating the proportion of simulation runs in which the null model of instantaneous speciation was rejected at the 5% level. Because the alternative hybridization model has two more parameters than the null model (if i is fixed to be 0.2), it is predicted that the distribution of logarithm of the likelihood ratio from the simulation under the null model, Δ , follows the χ^2 distribution with $df = 2$. We define Δ as the difference between the maximum log likelihood under the alternative hybridization model and that under the null model. We first checked this by performing 10,000 replication runs of coalescent simulations under the null model with the best-fit parameters (i.e., $\theta = 0.00405$ and $T_S = 1.20$). In each replication, 329,636 independent d values were simulated by using the ms software (Hudson 2002) to which our ML function was applied, and Δ was computed. The distribution of Δ in the simulation was slightly different from the χ^2 distribution; therefore, rather than using the χ^2 distribution, we decided to use the simulated distribution as the null distribution in which the 5% cutoff value was $\Delta = 2.17$. In other words, we treated Δ as a summary statistic and the 5% cutoff was determined by simulation. Similar approaches were used by Kim and Stephan (2002) and Kim and Nielsen (2004).

For the power analysis, we fixed $i = 0.2$ and 10,000 replication runs were carried out for each pair of T_H and T_I values. The power to detect hybridization was defined as the proportion of replications with $\Delta > 2.17$, which is denoted by $Q_{\Delta > 2.17}$. θ and T_S were set such that the average and SD of the simulated d were consistent with estimates obtained by analyzing the real data from the UCSC Genome Browser database (i.e., $\bar{d} = 0.0089$ and $SD = 0.0103$). Figure 4A summarizes the results of the power simulations when T_H was set equal to 0.01, 0.02, 0.05, or 0.1 with various values of T_I ($T_I = \{0.25, 0.5, 0.75, 1, 2\}$). The results for the four values of T_H were similar, indicating that our analysis was robust to the hybridization time T_H . As T_I increased, the power dramatically increased, suggesting that our ML approach had significant power to detect hybridization unless the time of isolation (T_I) was very short ($T_I \leq 0.5$).

FIG. 4. The power of the ML approach and its robustness to recombination and mutation rate variation, which are evaluated by the proportion of simulation runs with $\Delta > 2.17$. The arrow represents the parameter pair that is roughly in agreement with estimates for the human genome. See text for details.

Next, we investigated the effects of potential bias in our ML approach due to the factors ignored. We considered that the most problematic of these factors would be recombination within each 100-bp fragment (Wall 2003). Although our use of 100-bp fragments was designed to minimize the effect of recombination, it could not remove this effect completely. Another factor that was ignored was the effect of mutation rate variation; this effect may not have been negligible, but it should have been minimal because we used only non-CpG sites. We therefore performed simulations under the best-fit null model ($\theta = 0.00405$ and $T_S = 1.20$) with various levels of recombination and mutation rate variation. We considered three levels of population recombination rates, $\rho = \{0.5, 1, 2\} \times \theta$, because the recombination rate per bp would have been in the same order as the mutation rate (Bouffard et al. 1997; Nagaraja et al. 1997; Nielsen 2000). The mutation rate variation was incorporated into the model by assuming that the mutation rate for each fragment was a random variable from a gamma distribution. The shape of the gamma distribution was determined such that l was the coefficient of variation of the mutation rate (the SD of the mutation rate was l times the average mutation rate). The value of l was changed incrementally from 0.1 to 0.5 ($l = \{0.1, 0.2, 0.3, 0.5\}$). For all pairwise combinations of the recombination rate and l , we performed 10,000 replication runs of the simulations. In each replication, following the procedure for the power analysis described above, d for 329,636 independent fragments were simulated and the likelihood was computed. Then, $Q_{\Delta>2.17}$ was calculated as described above. The results are summarized in figure 4B.

We observed that $Q_{\Delta>2.17}$ tended to increase with increasing recombination rate, indicating that recombination (if ignored) could create false positive evidence for hybridization. One might consider this counterintuitive because recombination reduces the variance in the coalescent time, whereas temporal isolation increases it. Our results can be understood as follows. Because we assume no

recombination in our ML function, the null model will fit the data when the coalescent time has a sharp peak at τ_{species} , and the decay of such a sharp peak can be considered as evidence for temporal isolation. Recombination in a DNA region can reduce the variance in the coalescent times in that region, thereby causing a decay of the peak at τ_{species} . This is most likely the reason why recombination can create false positive evidence for hybridization.

The effect of the mutation rate variation is complicated because its behavior depends on the recombination rate. $Q_{\Delta>2.17}$ is highest at $l = 0.3$ when $\rho/\theta = \{1, 2\}$, whereas $Q_{\Delta>2.17}$ is highest at $l = 0.2$ when $\rho/\theta = 0.5$. Overall, the results show that unless $(\rho/\theta, l) = (0.5, 0.5)$, $Q_{\Delta>2.17}$ is approximately 0.05 or greater. Our findings suggest that, as long as $Q_{\Delta>2.17} \geq 0.05$, the joint action of recombination and mutation rate variation works to create false positive evidence for hybridization; therefore, our conclusion that there is no evidence for hybridization is robust.

From these simulation results, it may be reasonable to decide that our conclusion of no strong evidence for hybridization holds when taking into account the effect of recombination and mutation rate variation, given a reasonable range of $(\rho/\theta, l)$. There have been extensive researches on the rate of recombination in the human genome, and typical estimates range from 0.00038 to 0.0009 (Bouffard et al. 1997; Nielsen 2000; International Human Genome Sequencing Consortium 2001; Innan et al. 2003; Padhukasahasram et al. 2006). These values correspond roughly to 0.5–1.0 in ρ/θ . Unfortunately, to our best knowledge, a reliable estimate of l is not available. It is not very straightforward to estimate l from the human–chimpanzee divergence data set because of the substantial amounts of local variation in TMRCAs discussed previously. An exception is the nonrecombining Y chromosome in which TMRCAs for the entire chromosome is identical so that local variation in nucleotide divergence d directly reflects the mutation rate variation. Using Y chromosome divergence data obtained from

FIG. 5. Illustration of the data used in this study and Patterson et al. (2006).

the UCSC Genome Browser database, we have estimated $l = 0.20$. Therefore, if these estimates apply to the autosomes in humans and chimpanzees, we can conclude that there is no strong evidence for hybridization in the autosomes.

So far, we have shown the results with a fragment size of 100 bp. We chose this size because it would have reasonable power to detect hybridization while keeping the effects of recombination and mutation rate variation small. We verified our choice of 100 bp by simulations with fragments 50 and 200 bp long. As shown in the Supplementary Material online, if we used 50-bp fragments, there was a substantial reduction in the power to detect hybridization. If we used 200-bp fragments, the power increased; however, for all pairs of ρ and l values examined, the null model was rejected in more than $\sim 80\%$ of replications, indicating a very high level of false positives. Thus, even if we found evidence for hybridization by applying our ML equation (6) to 200-bp fragments, it would most likely be due to violation of the assumptions (i.e., the presence of recombination and mutation rate variation). Therefore, we consider it statistically reasonable to use 100-bp fragments rather than 50- or 200-bp fragments.

X Chromosome

Potential Bias in the HCGOM Data by Patterson et al. (2006)

First, we simply asked whether the data set of Patterson et al. (2006) satisfactorily represented the evolutionary pattern of the entire X chromosome. This question arises because their hybridization hypothesis relies strongly on observations made from the alignment sequence data of five primates, human, chimpanzee, gorilla, orangutan (O), and

macaque (M) (called the HCGOM data in Patterson et al. 2006; see also fig. 5) in which the authors found extremely small numbers of HG and CG sites. The HCGOM data analyzed by Patterson et al. (2006) covers only 0.24% of the X chromosome (~ 155 Mb). According to table 2 of Patterson et al. (2006), within the ~ 372 kb of aligned regions, there are 457 HC, 14 HG, and 12 CG sites, and the relative proportion of HG and CG sites is $\frac{n_{HG} + n_{CG}}{n_{HC} + n_{HG} + n_{CG}} = 0.054$. Here, we examined a larger data set, called the HCGM data set, which lacks orangutan sequences but covers more than twice as much of the X chromosome as is covered by the HCGOM data. In the HCGM data, the relative proportion of HG and CG sites is $\frac{95 + 86}{1035 + 95 + 86} = 0.149$, which is about three times that in the HCGOM data. This difference raises the possibility that there might be some kind of bias in the HCGOM data, most likely arising by chance because of the small amount of data.

We investigated what parts of the X chromosome are included in the HCGOM data. By following the screening process of the pairwise alignment data from the UCSC database described above, we obtained 112.9 Mb of aligned orthologous regions in the X chromosome (101.6 Mb after excluding CpG sites, segmental duplications, and tandem repeats); this covered about 72.8 % of the X chromosome. This data set was denoted $\mathcal{X}_{UCSC}^{ALL}$. We then scanned the $\mathcal{X}_{UCSC}^{ALL}$ data for regions that were also included in the HCGOM data. We successfully identified overlapping regions with total length 408 kb; these were denoted $\mathcal{X}_{UCSC}^{HCGOM}$. This length was in a good agreement with the total length of the aligned sequences in the HCGOM data (372 kb), indicating that we had successfully identified most, if not all, regions in the HCGOM data. The total length of sequences in our data set was slightly greater than that in the HCGOM data because the information only for variable sites is available in the data of Patterson et al. (2006). The supplemental data of Patterson et al. (2006) include a list of variable sites in the HCGOM data, so we extracted regions sandwiched by two divergent sites with no unusually long intervals. This method would have missed small regions that were excluded by Patterson et al. (2006) most likely because of low sequence quality.

The important point here is that the two data sets originated from an identical resource, i.e., $\mathcal{X}_{UCSC}^{HCGOM}$ was a subset of $\mathcal{X}_{UCSC}^{ALL}$. Therefore, the expected divergence in $\mathcal{X}_{UCSC}^{HCGOM}$ was identical to that in $\mathcal{X}_{UCSC}^{ALL}$. We found that the divergence at non-CpG sites was 0.00678 for $\mathcal{X}_{UCSC}^{HCGOM}$, which is $\sim 5\%$ smaller than the value of 0.00719 obtained for $\mathcal{X}_{UCSC}^{ALL}$ (table 1), but this difference might not have been statistically significant. We also performed the same analysis for the autosomes, and we found that the difference between the two data sets was smaller than that found for the X chromosome. The SD of d in $\mathcal{X}_{UCSC}^{HCGOM}$ was roughly 10% less than that in $\mathcal{X}_{UCSC}^{ALL}$. It was not clear how much of Patterson and colleagues' observation of very low divergence in the X chromosome could be explained by these relatively small differences. The answer will become apparent once the entire genome sequences for all five species become available.

Table 1. Summary of the Divergence in the Four Data Sets.

	Number of Fragments	$d (\times 10^{-3})$	
		Mean	SD
Autosomes			
$\lambda^{\text{ALL}}_{\text{UCSC}}$	329,636	8.90	10.3
$\lambda^{\text{HCGOM}}_{\text{UCSC}}$	19,471	8.70	9.89
X chromosome			
$\lambda^{\text{ALL}}_{\text{UCSC}}$	18,394	7.19	9.61
$\lambda^{\text{HCGOM}}_{\text{UCSC}}$	903	6.78	8.46

Discussion

Autosomes

We investigated the causes of discrepancy between two reports on the process of human–chimpanzee speciation. Innan and Watanabe (2006) showed that the pattern of divergence between the two species can be best explained by the simplest speciation model with instantaneous split, whereas Patterson et al. (2006) suggested that human–chimpanzee speciation involved temporal isolation followed by a hybridization event. Here, we focused primarily on the divergence in the autosomes because it is widely accepted that the autosomes are the best place to detect a footprint of a demographic event such as hybridization. The power to detect evidence of such past demographic events is maximized when genetic information from a huge number of independent autosomal loci is available (Innan and Watanabe 2006; Yang 2010). Patterson et al. (2006) proposed their complex speciation hypothesis on the basis of a large local variation of divergence time in the autosomes, but, as repeatedly pointed out (Barton 2006; Wakeley 2008; Presgraves and Yi 2009), they did not provide any statistical tests to support their hypothesis. The availability of the genome sequences of the two species allowed us to examine the hybridization hypothesis in a statistical framework.

We developed an ML function to detect hybridization by modifying the ML function of Innan and Watanabe (2006) and applying it to ~330,000 independent autosomal regions. We found that the data were best fit by the null model of instantaneous speciation (fig. 3), indicating that there was no evidence for hybridization in the autosomes. This was consistent with the conclusions of Innan and Watanabe (2006) but not those of Patterson et al. (2006). We also explored potential biases in our ML analysis due to recombination and mutation rate variation. Our additional simulations demonstrated that recombination and mutation rate variation were likely to produce false positive evidence for hybridization for a reasonable parameter range (fig. 4), so they had a conservative effect on our argument for no evidence for hybridization.

To further confirm our conclusion, we investigated the power of our ML approach to detect hybridization. Conditional on the amount of data used in this study, we found that our ML approach had substantial power to detect hybridization unless the time of isolation was very short ($T_1 \leq 0.5$). $T_1 = 0.5$ corresponds to 2.5 My if T_5 is assumed to be 6 My, indicating that our method had

reasonable power to detect a period of isolation longer than 2.5 My. The power would decrease when the time of hybridization was too long, but such ancient isolation is beyond the scope of this study. Our results led us to argue against Patterson and colleagues' hybridization hypothesis because the scenario they suggest implies quite a long time of isolation. Patterson et al. (2006) said, "if human and chimpanzee ancestors initially speciated and then interbred, hybrid males might have been infertile, consistent with Haldane's rule." They further suggested that "if a hybridization involving a single episode of gene flow occurred, it might result in a bimodal distribution of τ (local coalescent time);" this was listed as one of the "predictions that can be tested with larger data sets" (Patterson et al. 2006). To meet these statements, the two ancestral populations had to have accumulated some fixed incompatible alleles, which could cause hybrid sterility. The time needed for this should be at least in the order of $2N$ generations (or at least several Mys), and we showed that such a lengthy isolation period made the distribution of τ bimodal (fig. 2). Thus, we concluded that there was no strong evidence for evolutionarily meaningful isolation followed by hybridization in the human–chimpanzee ancestral population, although our result did not exclude the possibility of a very short temporal isolation or a very ancient isolation.

It should be noted that there is nothing conflicting between our results for autosomes and those of Patterson et al. (2006), except the interpretation of the analyses. Patterson et al. (2006) found that local coalescent time around HG and CG sites is on average 50% longer than that around HC sites. They suggested that this time interval should correspond to more than 4 My, reflecting a long period of isolation. In figure 1, however, we showed that Patterson and colleagues' observation is exactly what we expected under a simple model of instantaneous speciation. The coalescent time around HG and CG sites is long simply because of ascertainment bias: The coalescent process is prolonged because of our prior knowledge of the existence of a variable site.

In summary, the observations of Patterson et al. (2006) were well explained by the null model of instantaneous speciation with a large ancestral population. Taken together, our results suggested that there was no strong evidence to support the hybridization hypothesis. In a recent review (Siepel 2009), by examining the results of a number of studies that assessed the ancestral population size and speciation time (e.g., Ebersberger et al. 2007; Hobolth et al. 2007; Burgess and Yang 2008; Webster 2009), Siepel (2009) suggested that "these (Patterson et al. 2006) observations can be explained by realistically large ancestral population sizes, without the need to hypothesize a complex speciation event." Our results were consistent with this view.

X Chromosome

Analysis of the divergence of the X chromosome played a crucial role in the development of the hybridization hypothesis proposed by Patterson et al. (2006). They found that the human–chimpanzee divergence for the X chromosome is

much smaller than that for the autosomes, but no significant difference was observed for the human–gorilla comparison. This is a very surprising and puzzling observation, and it may not have a simple explanation. We did not include gorilla sequences in our analysis, and therefore, it was difficult to make a strong argument from our X chromosome analysis, but our results showed that the HCGOM data used by Patterson et al. (2006) were slightly biased toward low divergence regions (~5% reduction). Furthermore, the divergence in the HCGOM data used by Patterson et al. (2006) was less variable than that in our data set (table 1). However, these differences were not statistically significant and could only partially explain the surprising observation made by Patterson et al. (2006). Thus, although our analysis of autosomal data showed that there was no strong evidence for hybridization in the autosomes, for the reasons outlined above it is impossible to make a strong argument against the hybridization hypothesis from our analysis of the X chromosome.

Nevertheless, as has been extensively pointed out by recent reviews (Presgraves and Yi 2009; Siepel 2009), the observation made by Patterson et al. (2006) does not directly imply that hybridization has occurred because there are a number of factors that make interpretation of the observation by Patterson et al. (2006) difficult: for example, the effective population size of the X chromosome and male-biased mutation rate. A small effective population size would explain the observation of less variability in local coalescent time on the X chromosome, and a strong male-biased mutation rate would cause a reduction of the divergence on the X chromosome on average. There are conflicting arguments in regard to both these factors in primate evolution (for the X effective population size, Hammer et al. 2008; Bustamante and Ramachandran 2009; Keinan et al. 2009; for the male-biased mutation rate, Makova and Li 2002; Ellegren 2009; Presgraves and Yi 2009). Furthermore, because males are hemizygous for the X chromosome, the dominance effect of selection is quite different in the X chromosome compared with that in the autosomes (Charlesworth et al. 1987; Betancourt et al. 2004; McVicker et al. 2009). The point we emphasize here is that without determining the magnitude of the effect of these X-related factors during primate evolution it is not reasonable to make inferences concerning ancestral events on the basis of sex chromosomes. Acquisition of more genomic data in primates will facilitate our understanding of the human–chimpanzee speciation event (Siepel 2009). However, even when the genome sequences of all current primate species become available, there still would be technical limitations to the extraction of information on the evolutionary events that occurred several million years ago.

Acknowledgments

We are grateful to N. Patterson, M. Lascoux and anonymous reviewers for comments and to the members of the Innan Lab for discussions and technical helps. This work is supported from Japan Society for the Promotion of Science

(JSPS), Japan Science and Technology Agency, National Science Foundation, National Institute of Health and an internal grant of the Graduate University for Advanced Studies to H.I. M.Y. is a JSPS predoctoral fellow. J.G. is also supported by grants from JSPS.

References

- Barton NH. 2006. Evolutionary biology: how did the human species form? *Curr Biol*. 16:R647–R650.
- Becquet C, Patterson N, Stone AC, Przeworski M, Reich D. 2007. Genetic structure of chimpanzee populations. *PLoS Genet*. 3:e66.
- Benson G. 1999. Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Res*. 27:573–580.
- Betancourt AJ, Kim Y, Orr HA. 2004. A pseudohitchhiking model of X vs. autosomal diversity. *Genetics*. 168:2261–2269.
- Bouffard GG, Idol JR, Braden VV, et al. (14 co-authors). 1997. A physical map of human chromosome 7: an integrated YAC contig map with average STS spacing of 79 kb. *Genome Res*. 7:673–692.
- 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.
- Bustamante CD, Ramachandran S. 2009. Evaluating signatures of sex-specific processes in the human genome. *Nat Genet*. 41:8–10.
- Charlesworth B, Coyne JA, Barton NH. 1987. The relative rates of evolution of sex chromosomes and autosomes. *Am Nat*. 130:113–146.
- Chen FC, Li WH. 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.
- Cheng Z, Ventura M, She X, et al. (12 co-authors). 2005. A genome-wide comparison of recent chimpanzee and human segmental duplications. *Nature*. 437:88–93.
- Chimpanzee Sequencing and Analysis Consortium. 2005. Initial sequence of the chimpanzee genome and comparison with the human genome. *Nature*. 437:69–87.
- Coyne JA, Orr HA. 2004. Speciation. Sunderland (MA): Sinauer Associates Inc.
- Ebersberger I, Galgoczy P, Taudien S, Taenzer S, Platzer M, von Haeseler A. 2007. Mapping human genetic ancestry. *Mol Biol Evol*. 24:2266–2276.
- Ellegren H. 2009. The different levels of genetic diversity in sex chromosomes and autosomes. *Trends Genet*. 25:278–284.
- Hammer MF, Mendez FL, Cox MP, Woerner AE, Wall JD. 2008. Sex-biased evolutionary forces shape genomic patterns of human diversity. *PLoS Genet*. 4:e1000202.
- 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. Properties of a neutral allele model with intragenic recombination. *Theor Popul Biol*. 23:183–201.
- Hudson RR. 2002. Generating samples under a Wright-Fisher neutral model of genetic variation. *Bioinformatics*. 18:337–338.
- Huff CD, Xing J, Rogers AR, Witherspoon D, Jorde LB. 2010. Mobile elements reveal small population size in the ancient ancestors of *Homo sapiens*. *Proc Natl Acad Sci U S A*. 107:2147–2152.
- Innan H, Padhukasahasram B, Nordborg M. 2003. The pattern of polymorphism on human chromosome 21. *Genome Res*. 13:1158–1168.
- Innan H, Tajima F. 1997. The amounts of nucleotide variation within and between allelic classes and the reconstruction of the common ancestral sequence in a population. *Genetics*. 147:1431–1444.
- Innan H, Watanabe H. 2006. The effect of gene flow on the coalescent time in the human–chimpanzee ancestral population. *Mol Biol Evol*. 23:1040–1047.

- International Human Genome Sequencing Consortium. 2001. Initial sequencing and analysis of the human genome. *Nature* 409:860–921.
- Kaessmann H, Wiebe V, Pääbo S. 1999. Extensive nuclear DNA sequence diversity among chimpanzees. *Science* 286:1159–1162.
- Kaessmann H, Wiebe V, Weiss G, Pääbo S. 2001. Great ape DNA sequences reveal a reduced diversity and an expansion in humans. *Nat Genet.* 27:155–156.
- Keinan A, Mullikin JC, Patterson N, Reich D. 2009. Accelerated genetic drift on chromosome X during the human dispersal out of Africa. *Nat Genet.* 41:66–70.
- Kim Y, Nielsen R. 2004. Linkage disequilibrium as a signature of selective sweeps. *Genetics* 167:1513–1524.
- Kim Y, Stephan W. 2002. Detecting a local signature of genetic hitchhiking along a recombining chromosome. *Genetics* 160:765–777.
- Kimura M. 1969. The number of heterozygous nucleotide sites maintained in a finite population due to steady flux of mutations. *Genetics* 61:893–903.
- Kingman JFC. 1982. The coalescent. *Stoch Process Appl.* 13:235–248.
- Kitano T, Schwarz C, Nickel B, Pääbo S. 2003. Gene diversity patterns at 10 X-chromosomal loci in humans and chimpanzees. *Mol Biol Evol.* 20:1281–1289.
- Makova KD, Li WH. 2002. Strong male-driven evolution of DNA sequences in humans and apes. *Nature* 416:624–626.
- McVicker G, Gordon D, Davis C, Green P. 2009. Widespread genomic signatures of natural selection in hominid evolution. *PLoS Genet.* 5:e1000471.
- Nagaraja R, MacMillan S, Kere J, et al. (25 co-authors). 1997. X chromosome map at 75-kb STS resolution, revealing extremes of recombination and GC content. *Genome Res.* 7:210–222.
- Nielsen R. 2000. Estimation of population parameters and recombination rates from single nucleotide polymorphisms. *Genetics* 154:931–942.
- Padhukasahasram B, Wall JD, Marjoram P, Nordborg M. 2006. Estimating recombination rates from single-nucleotide polymorphisms using summary statistics. *Genetics* 174:1517–1528.
- Patterson N, Richter DJ, Gnerre S, Lander ES, Reich D. 2006. Genetic evidence for complex speciation of humans and chimpanzees. *Nature* 441:1103–1108.
- Presgraves DC, Yi SV. 2009. Doubts about complex speciation between humans and chimpanzees. *Trends Ecol Evol.* 24:533–540.
- Roach JC, Glusman G, Smit AFA, et al. (15 co-authors). 2010. Analysis of genetic inheritance in a family quartet by whole-genome sequencing. *Science* 328:636–639.
- Segmental Duplication Database [Internet]. Build 36/panTro2. Seattle (WA): University of Washington, Department of Genome Science, Eichler Lab. 2001- [updated 2006 Mar; cited 2011 Oct 11]. Available from: <http://humanparalogy.gs.washington.edu/>.
- Siepel A. 2009. Phylogenomics of primates and their ancestral populations. *Genome Res.* 19:1929–1941.
- Tajima F. 1983. Evolutionary relationship of DNA sequences in finite populations. *Genetics* 105:437–460.
- Takahasi KR, Innan H. 2008. Inferring the process of human-chimpanzee speciation. In: *Encyclopedia of Life Sciences (ELS)*. Chichester (UK): John Wiley & Sons.
- Takahata N, Satta Y. 1997. Evolution of the primate lineage leading to modern humans: phylogenetic and demographic inferences from DNA sequences. *Proc Natl Acad Sci USA.* 94:4811–4815.
- Takahata N, Satta Y, Klein J. 1995. Divergence time and population size in the lineage leading to modern humans. *Theor Popul Biol.* 48:198–221.
- Tandem Repeat Finder [Internet]. Version 4.04. Boston (MA): Boston University. 1999- [updated 2006 Mar 13; cited 2011 Oct 11]. Available from: <http://tandem.bu.edu/trf/trf.html>.
- UCSC Genome Browser Database [Internet]. 2003–2011. Santa Cruz (CA): University of California, Santa Cruz. [updated 2011 Oct 10; cited 2011 Oct 11]. Available from: <http://genome.ucsc.edu/>.
- Wakeley J. 2008. Complex speciation of humans and chimpanzees. *Nature* 452:E3–E4.
- Wall JD. 2003. Estimating ancestral population sizes and divergence times. *Genetics* 163:395–404.
- Webster MT. 2009. Patterns of autosomal divergence between the human and chimpanzee genomes support an allopatric model of speciation. *Gene* 443:70–75.
- Wiuf C, Donnelly P. 1999. Conditional genealogies and the age of a neutral mutant. *Theor Popul Biol.* 56:183–201.
- Yang Z. 2010. A likelihood ratio test of speciation with gene flow using genomic sequence data. *Genome Biol Evol.* 2:200–211.
- Yu N, Jensen-Seaman MI, Chemnick L, Ryder O, Li WH. 2004. Nucleotide diversity in gorillas. *Genetics* 166:1375–1383.