The Effect of Gene Flow on the Coalescent Time in the Human-Chimpanzee Ancestral Population

Hideki Innan* and Hidemi Watanabe†

*Human Genetics Center, School of Public Health, University of Texas Health Science Center at Houston; and †Graduate School of Information Science and Technology, Hokkaido University, Sapporo, Japan

The coalescent process in the human-chimpanzee ancestral population is investigated using a model, which incorporates a certain time period of gene flow during the speciation process. \( \alpha \) is a parameter to represent the degree and time of gene flow, and the model is identical to the null model with an instantaneous species split when \( \alpha = \infty \). A maximum likelihood (ML) method is developed to estimate \( \alpha \), and its power and reliability is investigated by coalescent simulations. The ML method is applied to nucleotide divergence data between human and chimpanzee. It is found that the null model with an instantaneous species split explains the data best, and no strong evidence for gene flow is detected. The result is discussed in the view of the mode of speciation. Another ML method is developed to estimate the male-female ratio (\( \xi \)) of mutation rate, in which the coalescent process in the ancestral population is taken into account.

Introduction

The mode of speciation has been one of the central themes in evolutionary biology (Mayr 1963). The simplest model of speciation assumes an instantaneous split of species, but this model may oversimplify a speciation event, because selection and gene flow could be involved during the process of speciation. For example, the genomic view of speciation describes this process such that speciation can be initiated by differential adaptation in small parts of the genome (i.e., “speciation genes,” see Wu and Ting [2004] for a recent review), where gene flow is restricted. On the other hand, gene flow is allowed in other parts of the genome. For a recent discussion on this issue, see Wu’s (2001) review with accompanying commentaries by various authors.

DNA sequence comparison between closely related species could be a powerful approach to resolve the mode of speciation. If speciation occurred instantaneously, the divergence between two descendant species can be explained by the simplest null model in which the ancestral population splits at a single time point. On the other hand, if not, genetic isolation in speciation genes and their flanking regions is formed earlier than other parts of the genome, making the variance of nucleotide divergence across the genome larger than that predicted under the simplest model. Therefore, it may be possible to distinguish these models of speciation by focusing on the variance of nucleotide divergence (Wakeley 1996). This approach is powerful especially for very closely related species that share many ancestral polymorphisms (shared polymorphisms) (Wakeley and Hey 1997). Successful applications of this method are seen in the Drosophila species (Wang, Wakeley, and Hey 1997; Machado et al. 2002).

The human-chimpanzee speciation event is of special interest. Because speciation is unfortunately too old to have many “shared polymorphisms,” polymorphism data are not as informative as the cases of Drosophila, and analysis should depend on divergence data between the two species. Initial studies by Takahata and colleagues (Takahata, Satta, and Klein 1995; Takahata and Satta 1997) showed that a maximum likelihood (ML) estimate of the “effective” size of the human-chimpanzee ancestral population may be several times larger than that of the modern humans. This large estimate may be surprising, but it is important to note that the effective size can be larger than the actual size if the ancestral population was structured and there was gene flow between them. The challenge is to test for gene flow from divergence data only.

The first purpose of this article is to extend Takahata’s ML method (Takahata, Satta, and Klein 1995) to investigate the effect of gene flow. Takahata’s method used divergence data from multiple nuclear regions (see also Takahata 1986). The average of human-chimpanzee nucleotide divergence may be around 0.012 per site in genome sequence data (Fujiyama et al. 2002), from which the divergence time is estimated to be roughly 6 MYA assuming a nucleotide mutation rate of \( 10^{-9} \) per year (e.g., Li 1997), in good agreement with phylogenetic estimation (e.g., 5.5 Myr [Kumar and Hedges 1998]).

Takahata’s method divides the divergence time into two parts: those before and after speciation (denoted by \( x \) and \( y \), respectively) assuming that speciation occurred at a single time point. The time before speciation is due to the coalescent process in the ancestral population. A recent application of this method to large-scale data (Satta et al. 2004) obtained estimates of \( x \) and \( y \) to be 0.0051 and 0.0073 (therefore, the sum is \( \sim 0.012 \)), respectively, suggesting that the ancestral population size may be several times larger than the effective population size of humans because estimates of nucleotide diversity in humans are around 0.0007–0.0008 (Hinds et al. 2005). These estimates of \( x \) and \( y \) are roughly in agreement not only with initial studies (Takahata, Satta, and Klein 1995; Takahata and Satta 1997) but also with recent ones with improved methods (Wall 2003), although Yang (2002) obtained a smaller estimate of \( x \) when the information of the shapes of gene trees is taken into account.

Here, we consider the model illustrated in figure 1 to incorporate gene flow during the period of speciation into the simple model with an instantaneous split. We assume that, since the beginning of speciation at time \( T_F \), the level of isolation increases linearly at rate \( \alpha \) and the complete isolation is established at time \( T_v \). This model is a special case of Teshima and Tajima (2002), and when \( \alpha = \infty \), the model is identical to Takahata’s null model. We develop an ML
Our general speciation model provides detailed information on the evolutionary process in the human-chimpanzee ancestral population. The second purpose of this article is to take this information into account in the estimation of the male-female ratio ($\alpha$) of mutation rate for humans, which has been controversial. A common way to estimate $\alpha$ is to compare the levels of divergence on autosomes and sex chromosomes. The most reliable estimate of $\alpha$ for humans should be obtained with the closest relatives (i.e., chimpanzees and bonobos), but such analysis is limited by the number of autosomes and sex chromosomes. The most reliable estimate in this article. It is assumed that the ancestral diploid population is initially panmictic and that the population size is constant, $N$. Time ($t$) is measured in units of $2N$ generations from the present ($t = 0$). Assume that the complete isolation between the two subpopulations, I and II, is established at time $t = T_S$, and they become different species, I and II, which correspond to humans and chimpanzees in this study. The population sizes of the two subpopulations, I and II, are assumed to be $hN$ and $(1 - h)N$, respectively. It should be noted that $h$ is a parameter that determines the ratio of the two subpopulations before $T_S$, and there is no assumption on the population sizes for $0 \leq t \leq T_S$, in which no coalescent event occurs as long as we consider only a pair of lineages from the two species.

Migration (gene flow) is allowed for $t > T_S$. Consider the process backward in time. It is assumed that the migration rate increases linearly with increasing $t$ and that the ancestral population eventually becomes panmictic with the effective size $N$. Let $M_{12}(t)$ and $M_{21}(t)$ be the population migration rates (population size, $N$, times migration rate per generation) from I to II and from II to I at time $t$. We assume $M(t) = hM_{12}(t) = (1 - h)M_{21}(t)$, which is given by

$$M(t) = a(t - T_S),$$

for $t > T_S$ and $M(t) = 0$ for $t \leq T_S$, when $a = \infty$, the model is identical to the simplest model with no gene flow.

Suppose that there are two sequences, one is from species I and the other from species II. The number of nucleotide differences between them is determined by the coalescent and migration events (pdf) of the coalescent time, $P(t|N, T_S, h, a)$ (Edwards and Beerli 2000; Arbogast et al. 2002; Rosenberg and Feldman 2002; Teshima and Tajima 2002). Consider the coalescent process of the ancestral lineages of the two sampled sequences backward in time. Let $Q_{11}$ be the state where there is one ancestral lineage in subpopulation I and the other is in subpopulation II. It is obvious that the state is $Q_{11}$ for $t < T_S$, so that the system essentially starts at $t = T_S$. For $t > T_S$, coalescent and migration events occur, and the state changes. There are three other possible states: $Q_{20}$ in which the two ancestral lineages are in subpopulation I, $Q_{02}$ in which the two ancestral lineages are in II, and $Q_1$ in which the two ancestral lineages have coalesced to their common ancestor. Let $q_{11}(t), q_{20}(t), q_{02}(t),$ and $q_1(t)$ be the probabilities that the state is $Q_{11}, Q_{20}, Q_{02},$ and $Q_1$ at time $t$, respectively. It is easy to obtain $q_{11}(t), q_{20}(t), q_{02}(t),$ and $q_1(t)$ numerically because they follow simple recursions:

$$q_{11}(t + \Delta t) = [1 - m_{12}(1 - m_{21}) - m_{21}(1 - m_{12})]q_{11}(t) + 2m_{12}(1 - m_{12})q_{20}(t) + 2m_{21}(1 - m_{21})q_{02}(t),$$

$$q_{20}(t + \Delta t) = [(1 - m_{12})^2 - \Delta t/h]q_{20}(t) + m_{12}(1 - m_{12})q_{11}(t) + m_{21}^2q_{02}(t),$$

$$q_{02}(t + \Delta t) = [(1 - m_{21})^2 - \Delta t/(1 - h)]q_{02}(t) + m_{12}(1 - m_{21})q_{11}(t) + m_{21}^2q_{20}(t),$$

$$q_1(t + \Delta t) = q_1(t) + [\Delta t/h]q_{20}(t) + [\Delta t/(1 - h)]q_{02}(t),$$

for a very small $\Delta t$, where $m_{12} = 2M_{12}(t)\Delta t$, $m_{21} = 2M_{21}(t)\Delta t$. The numerical solutions for these recursions are obtained with the initial conditions $q_{11}(T_S) = 1$ and

![Figure 1.—Illustration of the model.](image_url)
\( q_0(T_S) = q_0(T_S) = q_1(T_S) = 0 \). Then, we have the pdf of
the coalescent time as
\[
P(t|N, T_S, h, a) = \frac{[q_1(t + \delta t) - q_1(t)]/\delta t.}
\]

The calculation of the recursions should be continued until the population becomes completely panmictic \((t = T_P)\), and then the standard coalescent theory is applied for the case where the two sequences do not coalesce until \(T_P\). However, because this strict way is time consuming, the calculation can be terminated when the population is “quasi-panmictic.” This quasi-panmictic state can be defined such that the coalescent probability, \(P(t|N, T_S, h, a)\), is more than 99% of that predicted in a panmictic population with size \(N\). This method gives almost identical result to the strict method.

From equation (6), the pdf of the number of nucleotide differences \((d)\) is given by
\[
P(d|\theta, T_S, h, a) = \int_{\tau_3}^{\infty} \left( \frac{2\pi \tau}{d!} \right)^a \exp(-2\mu\tau) \frac{d\tau}{d!} P(\tau|N, T_S, h, a)d\tau,
\]
assuming a Poisson distribution for the number of mutations, where \(\theta = 4N\mu\) and \(\mu\) is the mutation rate per generation. This equation allows us to obtain ML estimates of the parameters \((\theta, T_S, h, a)\) given observed \(d\) from independent neutral regions. Note that equation (7) assumes no intragenic recombination. Because it may be extremely hard to incorporate recombination in this full-likelihood framework, a large number of very short regions are used to minimize the effect of recombination (see below and Discussion).

The ML method is applied to human-chimpanzee divergence data. From the random bacterial artificial chromosome end sequences of chimpanzees (Fujiyama et al. 2002), we extracted 39,520 regions with length \(\geq 100\) bp for which the quality of sequencing is high and the orthologous region in the human genome is unambiguously determined. This data set is denoted by \(D_1\) and used to calculate the average divergence (fig. 2). The average \(d\) for autosomes is 0.01237 from 38,754 regions (13 Mb in total), which is in agreement with other studies (Chen and Li 2001; Ebersberger et al. 2002). \(D_1\) is nearly identical to the data used by Satta et al. (2004). The average \(d\) for X chromosome is 0.00904 from 748 regions (250 kb in total), lower than that for autosomes, while the average \(d\) of Y chromosome is the highest, 0.01737 from 18 fragments (6.7 kb in total).

From each of the 38,754 autosomal regions, we randomly extracted a 100-bp fragment. After removing very few fragments with the number of nucleotide differences \((d)\) more than 10, we obtained the distribution of \(d\), which is denoted by \(D_A = \{0, 1, 2, \ldots, 10\}\), where \(\delta_i\) is the number of fragments with \(d = i\). In a similar way, the distribution of \(d\) for X chromosome \((D_X)\) is also obtained.

The log likelihood for \(D_A\) is computed using the following equation:
\[
\ln L(0, T_S, h, a|D_A) = \sum_{d=0}^{10} \delta_d \ln P(d|0, T_S, h, a, d \leq 10),
\]

where
\[
P(d|0, T_S, h, a, d \leq 10) = \frac{P(d|0, T_S, h, a)}{\sum_{i=0}^{10} P(i|0, T_S, h, a)}.
\]

In \(L(0, T_S, h, a|D_A)\) is numerically calculated as follows. First, we fix \(h\) and let the other three parameters vary freely. \(\theta\) and \(T_S\) are changed with increments of \(2 \times 10^{-6}\) and \(10^{-3}\), respectively, and \(\log_{10} a\) is changed with an increment of 0.2 up to 2, which is nearly identical to the case of \(a = \infty\). We use four values of \(h = 0.05, 0.1, 0.2,\) and 0.5.

Results

Detecting Gene Flow in the Human-Chimpanzee Ancestral Population

The ML method was applied to the human-chimpanzee divergence data. First, \(h = 0.2\) is fixed, which somehow reflects estimates of recent effective population sizes of the two species. The recent effective population size of chimpanzees may be several times larger than that of humans (Kaessmann, Wiebe, and Paabo 1999; Kaessmann et al. 2001; Kitano et al. 2003).

We found that the maximum \(\ln L = -57024.578\) is attained when \(a = 100\) (\(\log_{10} a = 2\), \(\theta = 0.005474\), and \(T_S = 1.225\). Figure 3A shows that the maximum \(\ln L\) conditional on \(a\) (profile likelihood) decreases with decreasing \(a\). The maximum \(\ln L\) under the simple model with an instantaneous species split \((a = \infty)\) is also calculated. When \(\theta = 0.005476\) and \(T_S = 1.250\), the maximum \(\ln L = -57024.576\) is obtained, which is slightly higher than that for \(a = 100\), indicating that the ML estimate of \(a\) may be very large, essentially \(\infty\). Figure 3B shows the observed distribution of \(d\) with the expectation under the simplest model with \(\theta = 0.005476\) and \(T_S = 1.250\). The two distributions are very similar, suggesting that the simplest null model explains the data very well.

Coalescent simulations (Hudson 2002) were carried out to examine the power and reliability of our ML approach. Under the simplest null model with \(\theta = 0.005476\) and \(T_S = 1.250\), 38,754 independent \(d\) values were simulated and the ML estimate of \(a\), \(\hat{a}\), was obtained. Based on 1,000 replications of this simulation, the distribution of \(\hat{a}\) is examined. As shown in figure 3A, the likelihood
curve of $a$ is nearly flat for $a \geq 10(\log_{10} a \geq 1)$, indicating that gene flow is so extensive when $a \geq 1$ that the coalescent process is very similar to the null model of instantaneous speciation. Therefore, we first consider the probability that $\hat{a} < 10(\log_{10} \hat{a} < 1)$. It is found that this probability is slightly less than 0.5 (fig. 4A), suggesting that simulated data under the null model support the null model with probability more than half under the null model.

Note that $\max[\ln L|a = 0.4]$ is about two units lower than $\max[\ln L]$, indicating that an approximate 95% confidence interval (CI) of $a$ is between 0.4($\log_{10} a = -0.4$) and $\infty$. This is roughly in agreement with the results of coalescent simulations under the null model, which demonstrate that the probability of $\hat{a} < 0.4$ is about 8% (fig. 4B). Therefore, we consider that our method could have reasonable power to detect this amount of gene flow ($a = 0.4$).

Next, we ask how much power our method has to detect gene flow when there is true gene flow. We simulated data under models with gene flow, and the distribution of ML estimates of $a$ was obtained for five different degrees of gene flow, $\log_{10} a = 0, -0.4, -0.6, -0.8, -1$, and $-1$. Figure 4 shows the two probabilities, $\text{Prob}(\log_{10} \hat{a} < 1)$ and $\text{Prob}(\log_{10} \hat{a} < -0.4)$. According to figure 3A, $\max[\ln L|\log_{10} a = 0]$ is about one unit lower than $\max[\ln L]$. For this amount of gene flow, the two probabilities are quite higher than those under the null model. When $\log_{10} a = -0.4$ for which $\max[\ln L] - \max[\ln L|\log_{10} a = -0.4] \approx 2$, $\text{Prob}(\log_{10} \hat{a} < 1)$ exceeds 0.8 and $\text{Prob}(\log_{10} \hat{a} < -0.4)$ is close to 0.5. $\max[\ln L|a]$ decreases drastically when $\log_{10} a$ is lower than 0.4 and the two probabilities increase. $\text{Prob}(\log_{10} \hat{a} < -0.4)$ exceeds 0.9 when $\log_{10} a = -1$.

It might be expected that the power to detect gene flow could be increased if longer fragments are used. However, as the length increases, the effect of recombination becomes more serious. Because our ML function (eq. 8) assumes no recombination, the effect of the violation of this assumption is investigated by simulation. Coalescent simulations (Hudson 2002) were carried out with three levels of recombination, and the distributions of $\hat{a}$ for simulated data were investigated. Figure 4 shows the probabilities of $\log_{10} \hat{a} < 1$ and $\log_{10} \hat{a} < -0.4$. If the population recombination rate per site ($\rho$) to be about half of the mutation rate ($\theta$) is assumed, both $\text{Prob}(\log_{10} \hat{a} < 1)$ and $\text{Prob}(\log_{10} \hat{a} < -0.4)$ are higher than those without recombination. This ratio is roughly consistent with the average of estimates of $\rho$ from polymorphism data on human chromosome 21 (Innan,
Padhukasahasram, and Nordborg 2003), although recombination rate should have substantial local variation (e.g., Crawford et al. 2004; McVean et al. 2004). If we use higher ratios of $p$ to $\theta$ ($p/\theta = 1$ and 2), larger $\text{Prob}(\log_{10} a < 1)$ and $\text{Prob}(\log_{10} a < -0.4)$ are observed. It is suggested that recombination might create false-positive evidence for gene flow. To minimize this effect of recombination, relatively short fragments (100 bp) are used in this study, but the effect of recombination on this fragment length may not be negligible.

Another possible violation of the assumptions in our model is mutation rate variation. The effect of the mutation rate variation is also investigated by simulations under the null model assuming that the mutation rate for each fragment is a random variable from a gamma distribution. The shape of a gamma distribution is determined such that the standard deviation of the mutation rate is $k$ times the average mutation rate, which is fixed to be $\theta = 0.005476$. As shown in figure 4A, as $k$ increases, $\text{Prob}(\log_{10} a < 1)$ increases, suggesting that the mutation rate variation could inflate the possibility to reject the null model. On the other hand, with increasing $k$, $\text{Prob}(\log_{10} a < -0.4)$ decreases. This is because, under our setting, almost all $\log_{10} a$ are $-0.2$ or $-0.4$ for a large $k$.

Thus, the violation of these two assumptions could create false-positive evidence for gene flow; therefore, the results of our ML method should be understood carefully especially when the method detects the signature of gene flow (i.e., lower ML estimate of $a$). However, there may not be significant effect on our application to the human-chimpanzee data because the simplest null model is best supported.

The analysis so far assumed $h = 0.2$, and we found that the simplest null model with instantaneous isolation ($a = \infty$) explains the data best. Very similar results are also obtained for $h = 0.05$, 0.1, and 0.5. As shown in figure 3A, $\text{max}[\ln L[a]]$ increases with increasing $a$, indicating that the best-fit model may be the simplest null model ($a = \infty$), in which $h$ does not matter. It is important to notice that $a$ and $h$ are nonidentifiable. That is, under our ML framework, it may not be possible to obtain point estimates of $a$ and $h$ simultaneously. For example, $\text{max}[\ln L[a]]$ for $h = 0.5$ and $\log_{10} a = -0.2$ is very similar to that for $h = 0.2$ and $\log_{10} a = -0.4$, and the expected distributions of coalescent time for the former is nearly identical to that for the latter, making it nearly impossible to distinguish the two cases. Because the expected distribution of $d$ is determined by the probabilities of coalescence and migration as shown in equation (6), $a$ and $h$ may be able to be summarized by a single parameter.

Male-Female Ratio of Mutation Rate

Our ML method demonstrated that the simplest null model with $\theta = 0.005476$ and $T_F = 1.250$ best explain the human-chimpanzee data under our framework. Here, conditional on this best-fit model, we estimate the male-female ratio ($\alpha$) of mutation rate. A moment estimate of $\alpha$ is easily obtained from the average of divergence on the X chromosome ($d_X = 0.00904$) in $D_T$. Let $X/A$ be the ratio of the neutral nucleotide substitution rate on the X chromosome to that on autosomes. If there is an equal sex ratio and no difference in the effect of selection between males and females, the ratio ($w$) of the effective population size for the X chromosome to that for autosomes is $w = 3/4$. Assuming this ratio, the expected $d$ for X is $0.005476(X/A)$ ($w + 1.250$). Therefore, a moment estimate of $X/A$ is 0.825, from which we have $\alpha = 3.22$ by solving $X/A = (2/3)/(2 + \alpha)/(\alpha + 1)$ (Miyata et al. 1987; Li 1997).

It is straightforward to obtain the log likelihood function of $X/A$ conditional on the best-fit model. Equation (8) is modified such that the effective size of ancestral population is reduced by a factor of $w = 3/4$ and the mutation rate is given by $\mu X/A$. The likelihood curve for the observed distribution of $d$ from 748 $\times$ 100–bp X-linked fragments ($D_T$) is shown by the solid curve in figure 5A. The ML estimate is given as $X/A = 0.79$ (approximate 95% CI = 0.73–0.85), from which the ML estimate of $\alpha = 4.4$ (approximate 95% CI = 2.6–9.5). Figure 5B shows that $D_T$ is in good agreement with the expected distribution given $\alpha = 4.4$.

An estimate of $\alpha$ strongly depends on $w$, the ratio of the effective population size for the X chromosome to that for autosomes. It has been pointed out that, because males are hemizygous for the X chromosome, selection could work in different ways in males and females (Charlesworth, Coyne, and Barton 1987; Shimmin et al. 1993), resulting in bias in the male-female ratio of the effective population size from $w = 3/4$ (Andolfatto 2001; Schaffner 2004). Aquadro,
Begun, and Kindahl (1994) suggested that, assuming other factors are equal, \( w \) could be lower than \( 3/4 \) under the hitchhiking model with continuous input of advantageous mutations (Maynard Smith and Haigh 1974; Kaplan, Hudson, and Langley 1989). On the other hand, \( w \) may be higher than \( 3/4 \) under the background selection model in which negative selection keeps eliminating deleterious mutations (B. Charlesworth, Morgan, and D. Charlesworth 1993; Charlesworth 1996). Begun and Whitney (2000) reported that nucleotide diversity on the X chromosome is significantly lower than \( 3/4 \) of that on an autosome in *Drosophila simulans*, while the difference in nucleotide divergence from *Drosophila melanogaster* is small.

For humans, Sachidanandam et al. (2001) reported that nucleotide diversity on the X chromosome is \( 59\% \) of that on autosomes. Here, we attempt to take this information into account in the estimation of the male-female mutation rate ratio. Because the male-female mutation rate ratio and the relative effective population size for the X chromosome \( (w) \) could contribute the low nucleotide diversity on the X chromosome (e.g., Schaffner 2004), we have \( X/A \times w = 0.59 \). If it is assumed that \( w \) is the same in the human population and the ancestral population, \( X/A \times w = 0.59 \) also holds in the ancestral population. Then, by solving \( 0.005476(X/A)[0.59/(X/A) + 1.250] = 0.00904 \), a moment estimate of \( X/A \) is obtained to be 0.848. From this estimate, \( x \) is estimated to be 2.7, which is slightly lower than those assuming \( w = 3/4 \).

**Discussion**

**Detecting Gene Flow**

An ML method is developed to estimate the level of gene flow \( (a) \) together with the ancestral population mutation parameter \( (\theta) \) and speciation time \( (T_S) \) from a large number of independent autosomal regions. The population model includes a certain time period of gene flow between the ancestral populations of humans and chimpanzees. The larger the \( a \) the shorter the period of gene flow, and when \( a = \infty \), the model is identical to the simplest null model with no gene flow (Takahata 1986; Takahata, Satta, and Klein 1995; Takahata and Satta 1997).

The method is applied to the human-chimpanzee divergence data from 38,754 essentially independent autosomal regions. The ML estimates of the three parameters are \( a = 0.005476, \theta = 0.0005447 \), and \( T_S = 1.250 \), indicating that the simplest null model with no gene flow best explains the data. Under the null model of Takahata and colleagues (Takahata, Satta, and Klein 1995; Takahata and Satta 1997), our estimates translate \( x = 0 \) and \( y = T_S \theta \), which are in good agreement with other studies (Takahata, Satta, and Klein 1995; Takahata and Satta 1997; Chen and Li 2001; Ebersberger et al. 2002; Wall 2003; Satta et al. 2004) (but see Yang [2002] for a smaller estimate of \( x \)).

Because the regions are mostly from noncoding sequences, our estimate could be less affected by selection. Another advantage of our estimate is that the effect of recombination is small, which could cause the underestimation of \( \theta \) and the overestimation \( T_S \). The effect of recombination is minimized by using a large number of short fragments in this study. Our estimate of \( \theta \) is about 10% larger than that of Satta et al. (2004), who used nearly identical data but the lengths of regions are longer.

Very recently, a draft sequence of the chimpanzee genome was released (Chimpanzee Sequencing and Analysis Consortium 2005). Over 170,000 orthologous fragments (100 bp each) were collected across the autosomal regions in this larger data set, and our ML method is applied. Almost identical result is obtained: the simplest null model is best supported, although an approximate 95% CI is narrowed down to \( a > 1.58 \) \((\log_{10} a > 0.2)\), when \( h = 0.2 \).

Our ML approach introduced in this article has several limitations. First, intralocus recombination is ignored. Because it is very hard to incorporate recombination into the likelihood function, rather we assume no recombination in the ML framework, and the effect of recombination was investigated by simulation. We found that recombination has the tendency to create false-positive evidence for gene flow, although we used relatively short fragments (100 bp) to minimize the effect of recombination. Second, the changes of gene flow rate in the ancestral population could be more complicated than our model with a linear function. However, it is straightforward to use any function of gene flow in the recursion equations (2)–(5). Note that these problems may not affect our result that the null model with no gene flow explains the data best. Third, we assume a constant size of the ancestral population, but changes in ancestral population size affect the distribution of the coalescent time. Unfortunately, it may be extremely hard to estimate the changes of the ancestral population size. To relax these limitations, more complicated models have to be used. For such models, it may not be easy to obtain analytical expressions for the likelihood of the data. Probably, it is more feasible to rely on simulation-based methods for the evaluation of (approximate) likelihood (e.g., Marjoram et al. 2003; Wall 2003; Hey and Nielsen 2004), although computationally intensive.

**Are There Speciation Genes?**

Our ML approach failed to reject the null model with an instantaneous species split for the human-chimpanzee speciation. This result does not disagree with the recently proposed hypotheses on the human-chimpanzee speciation event (Navarro and Barton 2003b; Osada and Wu 2005). Osada and Wu (2005) compared the variance of the human-chimpanzee sequence divergences in coding regions and that in noncoding regions. The rationale behind this analysis is that the former could be larger if some of the coding regions were involved in speciation so that the divergences in such regions are much larger than others. About 98%–99% of our genome-wide random fragments are noncoding regions, from which we did not find strong evidence for gene flow. It might be indicated that noncoding regions could be a good control to compare with regions of interests, for example, coding regions in the approach of Osada and Wu (2005).

The chromosomal speciation model of Navarro and Barton (2003a, 2003b) is a special version of the above model of speciation genes except that very long regions (i.e., inversions) are considered as speciation genes because of the suppression of recombination between different
chromosomal types (Rieseberg 2001; Hey 2003). It might be expected that speciation inversions could be more detectable than speciation genes. Navarro and Barton (2003b) found more genes that might have been under adaptive selection on chromosomes which have differences in chromosomal rearrangements between humans and chimpanzees (i.e., human chromosomes 1, 2, 4, 5, 9, 12, 15, 16, 17, and 18) than on the other colinear autosomes, suggesting that the rearranged chromosomes might have played an important role in the speciation event (although reanalysis of Lu, Li, and Wu [2003] found no significant difference).

This hypothesis also predicts elevated levels of silent divergence on rearranged chromosomes, but the authors did not observe this trend (see also Zhang, Wang, and Podlaha 2004). Our larger data set again found very similar levels of divergence: 0.01226 for rearranged and 0.01238 for colinear chromosomes. The distributions of d for the two classes of chromosomes are extremely similar (data not shown). However, comparing two groups of chromosomes may not have sufficient power to detect speciation inversions if a small number of inversions are involved in speciation. More powerful approaches for testing hypotheses and fitting models are needed to understand the role of inversion in speciation events, for example, computational intensive approaches as discussed in the previous section.

Male-Female Ratio of Mutation Rate

Our ML estimate of $\alpha$ (2.7–4.4) from the human-chimpanzee comparison is close to previous estimates ($\alpha \approx 5$) from comparisons of distantly related primates (Li 1997; Makova and Li 2002), rather than low estimates from human-chimpanzee sequence data (Bohossian, Skaletsky, and Page 2000; International Human Genome Sequencing Consortium 2001), which were obtained by ignoring the effect of the coalescent in the ancestral population. It is indicated that taking the effect of the coalescent in the ancestral population into account is important in estimating $\alpha$.

There are caveats for our estimates of $\alpha$. $\alpha$ was first estimated assuming $w = 3/4$ and equal rate of nucleotide substitution for autosomes and the X chromosome. However, because males are hemizygous for the X chromosome, selection could work in different ways between males and females (Charlesworth, Coyne, and Barton 1987). This could result in the violation of the two assumptions unless all mutations are neutral (Shimmin et al. 1993; Andolfatto 2001; Schaffner 2004). The effect of selection on $w$ was somehow adjusted by plugging the observed ratio of the level of polymorphism on X to autosomes in humans into point estimation, but the problem on the nucleotide substitution rate difference between X and autosome after speciation still remains. These are among the many open questions in population genetics and molecular evolution of sex chromosomes.

Acknowledgments

We thank N. Osada, N. Rosenberg, N. Saitou, K. M. Teshima, J. Wall, and an anonymous reviewer for comments. H.I. is supported by grants from the University of Texas and H.W. is partially supported by JSPS core-to-core program HOPE and the Ministry of Education, Culture, Sports, Science, and Technology Grant-in-Aid for Scientific Research on Priority Areas Comparative Genomics, 17018002, 2005, and Scientific Research (C), 15510158, 2005.

Literature Cited


Naruya Saitou, Associate Editor

Accepted February 13, 2006