ESTIMATING THE PROCESS OF SPECIATION FOR HUMANS AND CHIMPANZEES

By

YONG WANG

A Dissertation submitted to the

Graduate School-New Brunswick

Rutgers, The State University of New Jersey

and

The Graduate School of Biomedical Sciences

University of Medicine and Dentistry of New Jersey

in partial fulfillment of the requirements

for the degree of

Doctor of Philosophy

Graduate Program in Microbiology and Molecular Genetics

written under the direction of

Jody Hey

and approved by

_____

_____

_____

_____

New Brunswick, New Jersey

October, 2009

ABSTRACT OF THE DISSERTATION

Estimating the process of speciation for humans and chimpanzees

By YONG WANG

Dissertation Director:

Jody Hey

One of the most fascinating questions for evolutionary scientists is “How did humans arise as a new species?” In the last seventy years, two major schools of theory, allopatric speciation and sympatric speciation, have been developed and applied to explain the speciation process. Allopatric theory attributes the inducement of speciation to the establishment of geographic barriers that abruptly divide the ancestral population into two reproductively isolated groups, while sympatric theory emphasizes the role of divergent selection, leading to assortive mating and gradually diminishing gene flow. The two different scenarios should leave distinct footprints in the derivative genomes of the emerging species. Many mathematical methods have been developed to study human-chimpanzee speciation history by studying the genetic variation pattern in current human and chimpanzee populations. However, most methods either fail to incorporate sympatric speciation, or use datasets that don’t provide enough information about ancient divergence. In this study, we developed a new maximum likelihood method for analyzing genome data under the ‘isolation with migration’ model. Testing with simulated datasets demonstrates that this method is capable of generating accurate estimates regarding both current and ancient evolutionary histories. We applied this

method to the whole-genome alignment of human, chimpanzee and orangutan. The estimated human-chimpanzee speciation time is 4.3 million years (Myr). This estimate is in agreement with several previous studies. A more important finding of our study is a weak but significant one-way gene flow from the chimpanzee to the human population (0.002 migrations per generation). Simulation studies confirm that this gene flow is not an artifact created by within-locus recombination or violation of other assumptions of our method. A further analysis finds that the gene flow from chimpanzees into humans and chimpanzees persisted for a limited period of time, subsequent to the initial separation. These results lead us to favor a speciation process for humans and chimpanzees that includes some limited genetic exchange.

Dedication

To my parents

Acknowledgement

First and foremost, I would like to express my deepest gratitude to my advisor, Dr. Jody Hey, who has supported me throughout my academic program with his patience and knowledge. Without his guidance and encouragement this thesis would not have been completed.

Next, my sincere thanks to the rest of my thesis committee: Dr. Peter Smouse, Dr. Tara Matise and Dr. David Madigan, for their insightful comments and inspirational suggestions. In particular, I would like to thank Dr. Peter Smouse for reading the first draft of this thesis and offering many valuable opinions.

I am also heartily thankful to my fellow labmates in the Hey lab: Yong-Jin Won, Makoto Shimada, Alivia Dey, Sang Chul Choi and Jungwoo Jung, for the stimulating discussions and for all the fun we have had in the last six years.

And last of all, I would like to thank my parents Jiaquan Wang and Meijuan Shi for their love and support.

Table of Contents

Lists of Tables

List of Illustrations

Chapter One

Introduction

Speciation is an evolutionary process that splits a single species into two reproductively isolated lineages, and understanding the speciation process is fundamental to evolution research. A classic model for speciation process was described by Dobzhansky (1936) and Muller (1940). In their model, the ancestral species was separated by a geographic barrier (major mountains, rivers, etc.) that completely blocked gene flow between two isolated (allopatric) subpopulations. Independent evolution in the two subpopulations led to accumulation of randomly fixed incompatible mutations, which then created reproductive isolation and finalized the speciation process. This model is called allopatric speciation. An alternative model for this process is sympatric speciation (Maynard Smith, 1966) (Felsenstein, 1981). The sympatric speciation model does not invoke a geographic separation, but rather suggests that divergent selection, driven by competition for resources, habitats and mates, creates local adaptation within a spatially homogenous population. The local adaptation then induces assortive mating via pleiotropy (and/or hitchhiking) and reproductive isolation gradually develops between adaptively divergent subpopulations. A major difference between the two speciation models concerns the question of whether gene flow exists during the process of speciation. Because gene flow has a homogenizing effect that works against divergence, it is believed that strong divergent selection is required to overcome the effect of gene flow and allow speciation. As a result, Mayer {Futuyma, 1980 #126} suggested that one should use allopatric speciation as a null model when study the speciation process.

Over the last several decades, many experiments have been designed to duplicate the process of speciation in particular organisms (mostly fruit flies). In their 1993 review,

Rice and Hostert (Rice and Hostert, 1993) summarized the results from a large number of experimental studies. They found no conclusive evidence for genetic drift to create reproductive isolation among isolated populations. On the other hand, multiple studies provided strong support for the establishment of reproductive isolation under divergent selection in allopatric populations. Using strong and multifarious divergent selection, several studies were able to find the same isolation in sympatric populations. These results lend feasibility to sympatric speciation theory. To accompany the experimental results with theoretical support, Rice and Hostert proposed a scenario where moderately strong divergent selection initially creates partial reproductive isolation. As this partial reproductive isolation reduces gene flow, other characters can become more gently selected, generating further isolation. After a number of cycles, this positive-feedback, run-away process will ultimately proceed to the level of complete reproductive isolation.

The human-chimpanzee speciation event is one of the more interesting objects in evolutionary research. “When did humans separate from chimpanzees?” “Which evolutionary forces drove this process?” These are important questions to be answered by evolutionary biologists. Although it is impossible to duplicate this speciation process in the laboratory, historical evolution has left distinct footprints in current human and chimpanzee genomes, which allow us to infer their evolutionary history by studying the genetic variation pattern found in current human and chimpanzee populations. Early studies of human-chimpanzee speciation have been dominated by an assumption of allopatric speciation. According to the allopatric model, the human-chimpanzee divergence consists of two components, divergence both before and after the speciation

event, proportional to the ancestral population sizes and the speciation time respectively (Takahata, 1986). Based on this idea, Takahata and colleagues studied 13 pairs of human and chimpanzee sequences and obtained the maximum likelihood estimates of the two divergence components (0.005 prior to speciation, 0.010 after speciation). From these two estimates, they calculated that humans diverged from chimpanzees approximately 4.6 Myr ago and the effective population size of the human lineage was ~83,000, before speciation.

Another widely used method for studying human evolutionary history is called the tree mismatch method (Nei, 1987) (Wu, 1991), which exploits the fact that ancestral polymorphism creates conflicts between the species tree and the gene tree and estimates the ancestral population size, along with the speciation time, by equating the proportion of mismatched gene trees to the theoretical expectation (Rannala and Yang, 2003). Chen and Li (Chen and Li, 2001) applied this method to 53 coding contigs from human, chimpanzee, gorilla and orangutan, and obtained estimates of 6.2 and 8.4 Myr for human-chimpanzee and human-gorilla speciation time, with an estimate of 96,000 for the effective population size of the common ancestor of humans and chimpanzees.

In addition, many likelihood-based methods have been developed on the basis of Felsenstein's classic treatment (Felsenstein, 1988), which relates a dataset to the parameters of a population model by introducing the unknown gene tree (G) as a nuisance variable that is removed by integration (Hey, 2006).

$$L(Parameters \mid Data) = \sum_{G} \Pr(Data \mid G) \Pr(G \mid Parameters) \quad (1.1)$$

Burgess and Yang (Burgess and Yang, 2008) developed a method that approximates the integration in (1.1) using a Markov Chain Monte Carlo simulation. Their method was then used for analysis of a large data set of ~7.4 Mb aligned sequences from 5 primate species (Patterson *et al.*, 2006). The results showed that human and chimpanzee populations diverged from each other at about 4 Myr ago.

In recent years, evolutionary biologists have begun to look at the human-chimpanzee speciation process from a sympatric point of view. If it is really selection that gradually leads to complete reproductive isolation, genetic exchange should then be prohibited first in regions surrounding genes under strong divergent selection (or called by the name "speciation genes"). As a result, these "speciation genes" should diverge much earlier than the rest of the genome and demonstrate greater divergence, relative to the genomic average (Takahasi and Innan, 2008). Based on this idea, Osada and Wu (Osada and Wu, 2005) adapted Takahata's approach (Takahata, 1986) to compare human-chimpanzee divergence in coding versus non-coding sequences. Their likelihood ratio test revealed a significant smaller average divergence in non-coding sequences. This result matched the expectation of sympatric speciation, as coding sequences are more likely to be subject to selection. In another study, Navarro and Barton (2003) examined the idea that chromosomal rearrangement served the role of "speciation genes" (Rieseberg, 2001). They analyzed 115 genes (59 from nine pericentric inversions and one choromosomal fusion, 56 from co-linear chromosome sequences) in terms of the non-synonymous versus synonymous substitution ratio (K_A/K_S). They found that the ratio in rearranged

chromosomes was more than twice as large as that in co-linear chromosomes, supporting the role of chromosomal rearrangements in promoting the reproductive isolation. However, their conclusion has been criticized by other authors (Lu *et al.*, 2003) (Hey, 2003) and should be viewed with caution. Another controversial study by Patterson (Patterson *et al.*, 2006) compared human-chimpanzee divergence between sequences on the X chromosome and those on the autosomes. Their results indicated reduced divergence along the entire X chromosome, even after accounting for the difference in population size and mutation rates between X chromosome and the autosomes. To explain the phenomenon, Patterson proposed a hypothesis that a period of hybridization happened, following the initial isolation of human and chimpanzee populations. A flaw of Patterson's conclusion, as pointed out by Barton (Barton, 2006) and Wakeley (Wakeley, 2008), is that he did not statistically test his hypothesis against the null hypothesis of simple allopatric speciation. In addition, Wakeley (Wakeley, 2008) found that Patterson used a relatively small male-to-female mutation ratio, which will result in the underestimation of X chromosome divergence . Several attempts have been also been made to extend the likelihood method based on Felsenstein's equation to study the gene flow during divergence process. For example, Innan and Watanabe developed a maximum likelihood method to detect gene flow in the ancestral population (Innan and Watanabe, 2006). Applying their method to a dataset of ~17 Mb of human-chimpanzee orthologs, they found no evidence for gene flow. However, they did not rule out the possibility of sympatric speciation, though they did suggest that using a larger dataset would help to infer the true evolutionary history.

Many methods described above are intended for data sets with samples from many individuals at a few loci. One difficulty for these methods is, because the effective population size of modern humans and chimpanzees has been relatively small in the recent past, almost all lineages within each species will have coalesced more recently than the human-chimpanzee speciation event. Therefore, sampling more individuals at known positions will not provide much extra information (Barton, 2006). Fortunately, with recent improvements in DNA sequencing methodology, entire-genome sequences have become available from at least small numbers of humans, chimpanzees and other apes. Using rapid methods to analyze these new data, we should be able to gain some new insight into the speciation process for humans and chimpanzees.

The current research is aimed at studying the process of human-chimpanzee speciation. In particular, it aims to examine whether gene flow existed during the speciation process. Chapter Two describes a newly developed maximum likelihood method for analyzing genomic data under the 'isolation with migration' model. Unlike many other coalescent-based likelihood methods, this method does not rely on sampling genealogies, but rather provides a precise calculation of the likelihood by numerical integration over all genealogies. Simulation studies demonstrate that this method generates accurate estimates for population parameters, including gene migration rates, and therefore has the statistical power to differentiate between allopatric and sympatric speciation models. The maximum-likelihood method was then applied to the whole-genome alignment of human, chimpanzee and orangutan. Chapter Three presents the results of the estimation. Significant one-way gene flow from chimpanzees to humans

(0.002 migrants per generation) is reported. Results from simulation studies, which confirm that gene flow is not an artifact created by recombination or violation of other assumptions, are also presented. A further analysis finds that gene flow is likely to be restricted to a time period following initial separation. These findings are in agreement with the expectation of sympatric speciation theory. The final chapter concludes our current research and presents the significance of our results. Some limitations of our study are also discussed.

Chapter Two

Estimating Divergence Parameters with Small Samples from a Large Number of Loci

ABSTRACT

Many methods have been developed that adapt coalescent models to the divergence process between closely-related populations. Most methods are intended for data sets with samples from many individuals at few loci. Such data are good for inferring recent population history but are unlikely to contain much information about more ancient divergence. In recent years, the growing availability of genome sequences offers another potential source of data. Data sets extracted from whole genome alignments include DNA sequences sampled from very few individuals but at a very large number of loci. To take advantage of these data, we developed a new maximum likelihood method for analyzing genomic data under the ‘isolation with migration' model. Unlike many coalescent-based likelihood methods, the method does not rely on sampling genealogies, but rather provides a precise calculation of the likelihood by numerical integration over all genealogies. We demonstrate that our method works well on simulated data sets. We also consider two models for accommodating mutation rate variation among loci. We find the model that treats mutation rates as random variables leads to better estimates.

INTRODUCTION

In the study of speciation researchers often inquire of the extent that populations have exchanged genes as they diverged, and on the time since populations began to diverge. Answers to questions about historical divergence and gene flow potentially lie in patterns of genetic variation that are found in present day populations. To bridge the gap between population history and current genetic data, population geneticists often make use of a gene genealogy G as a nuisance variable (Griffiths, 1989). A gene genealogy is a bifurcating tree that represents the history of ancestry of sampled gene copies. The introduction of G provides a way to connect parameters of a model of divergence to the data. The probability of a particular value of G can be calculated for a particular parameter set, typically using coalescent models. Then given a particular genealogy, genetic variation can be examined using a mutation model that is appropriate for the kind of data being used. Finally by considering multiple values of G, the connection can be made between the population evolution history and the data. A mathematical representation that treats G as a key interstitial variable was given by Felsenstein (Felsenstein, 1988):

$$L(\Theta \mid X) = \Pr(X \mid \Theta) = \int_{\Psi} \Pr(X \mid G) \Pr(G \mid \Theta) dG , \qquad (2.1)$$

where X represents the sequence data, G represents gene genealogy, Ψ represents the set of all possible genealogies and Θ represents the vector of population parameters included in the model.

Unless sample sizes are very small (2.1) cannot be solved analytically, and so considerable effort has gone into finding approximate solutions (Griffiths, 1989; Kuhner *et al.*, 1995; Wilson and Balding, 1998). One general approach is to sample genealogies using a Markov chain Monte Carlo simulation. Kuhner and colleagues (Kuhner *et al.*, 1995) used an MCMC simulation to draw a large sample of genealogies from the distribution $\Pr(G|X, \Theta_0)$, conditioned on a driven parameter vector Θ_0. With a sample of genealogies the relative likelihoods for other values of Θ are evaluated by importance sampling. Another approach (Nielsen, 2000; Rannala and Yang, 2003; Wilson and Balding, 1998) uses a prior probability, $\Pr(\Theta)$, and MCMC simulations are performed to sample (G_i, Θ_i) pairs from the joint posterior distribution $\Pr(G, \Theta|X)$. Given enough running time, the marginal density curve of Θ_i will converge to $\Pr(\Theta|X)$. Both approaches allow extensions to the basic coalescent to include migration (Beerli and Felsenstein, 1999, 2001; Nielsen and Wakeley, 2001). A general problem for these methods is that they usually require long running times to generate sufficiently large and independent samples, especially when the MCMC simulation is mixing slowly.

With fast-improving DNA sequencing techniques, more and more genome sequences are becoming available, and alignments of these whole-genome sequences provide a potentially very useful source of information for the study of divergence. However traditional MCMC methods are likely to be slow on genome-scale data because running times are proportional to the number of loci. To overcome this difficulty Yang developed a likelihood method (Yang, 2002) for data sets containing one sample from each of the three populations at every locus. This method uses numerical integration to

calculate the likelihood function in formula (2.1). By using a very large number of loci, the method can make up for using a very small number of individuals (i.e. genomes).

Yang's method is based on a divergence model which assumes no gene flow between separated populations. However there are many situations where gene flow may have been occurring and where it is preferable to use a model that includes gene flow (Hey, 2006) (Nosil, 2008). One model that has been used frequently in this context is the 'isolation with migration' (IM) model, which incorporates both population separation and speciation and migration (Nielsen and Wakeley, 2001).

However, it is not straightforward to extend a numerical integration method, such as Yang's (Yang, 2002), to a model that includes gene flow. Under an IM model the genealogies include not only some fixed number of coalescent events and speciation events, but also any possible number of migration events. The potential for very large numbers of migration events complicates the sample space of G and makes the numerical integration seemingly impossible. Innan and Watanabe (Innan and Watanabe, 2006) circumvent this problem by using a recursion method to estimate the coalescent rates on a series of time points. In recursion, the accuracy in calculating coalescent rate at one time point depends on the accuracy of calculation at previous time points. This may impair the precision of the likelihood calculation. Therefore, we have developed a method that solely relies on numerical integration to calculate the likelihood under an IM model.

THEORY AND METHODS

We employ a two-population IM model (Figure 2.1) and assume selective neutrality. For convenience the two extant populations and the ancestral population are named Pop1, Pop2 and PopA respectively. For any one population the population size parameter is $\theta=4Nu$, where N is the effective population size and u is the neutral substitution rate. The population size parameters for the three populations in the model are denoted as θ_1, θ_2 and θ_A. A migration event from Pop1 to Pop2 (in the coalescent direction, back in time) is represented by $M_{1\rightarrow2}$ and a migration event in the reverse direction is represented by $M_{2\rightarrow1}$. Migration rate parameters have units of migrations per mutation event, i.e. $m=m/u$, where m is the migration rate per generation. Rates of the two kinds of migration events are denoted as m_1 and m_2. The speciation time parameter is $T=tu$, where t is the time since splitting in generations. In total the model includes six parameters: θ_1, θ_2, θ_A, m_1, m_2 and T.

One key to integrating over genealogies with migration events is to realize that the probability of the data given the genealogy is unaffected by migration events in the genealogy; $\Pr(X|G)$, depends on G only through branching topology and branch lengths. In other words, all genealogies that share the same coalescent events contribute identically to $\Pr(X|G)$. Let G^* denote a group of genealogies with the same coalescent events (but different migration events). If there is a way to calculate $\Pr(G^*|\Theta)$ together for all genealogies in G^*, then

$$L(\Theta \mid X) = \Pr(X \mid \Theta) = \int_{\psi^*} \Pr(X \mid G^*) \Pr(G^* \mid \Theta) dG^* \qquad (2.2)$$

The new integrand is estimated over the sample space of G^*, which is of much lower dimensionality relative to G. Here, we show that for the simple case where only a pair of genes are sampled from two populations, $\Pr(G^*|\Theta)$ can be calculated directly for the 'isolation with migration' model. The performance of the method is tested on simulated data sets.

Coalescent time distribution

Two gene copies are sampled at each locus, and we consider first the case when one is from Pop1 and the other from Pop2. These two genes coalesce at some time point t. If the coalescent event happened before both genes enter the ancestral population (i.e. $t<T$), then an odd number ($2x+1$, $x=0, 1, 2\ldots$) of migration events must occur before they coalesce, dividing t into $2x+2$ time intervals. During each interval, the ancestral lineages of the two samples reside in one of the three possible states:

$\mathbf{S_{11}}$: both ancestral lineages are in Pop1,

$\mathbf{S_{12}}$: one ancestral lineage is in Pop1 and the other is in Pop2,

$\mathbf{S_{22}}$: both ancestral lineages are in Pop2.

A migration event will result in a specific switch from one state to another, as shown in Figure 2.2. A coalescent can only happen in two states ($\mathbf{S_{11}}$ or $\mathbf{S_{22}}$), when both genes are in the same population. Assuming they coalesce in $\mathbf{S_{22}}$, there have been $2x+1$ migration events, $x+1$ of which are $M_{1\to2}$ and x of which are $M_{2\to1}$. Furthermore, of the $2x+2$ time intervals: $x+1$ are in state $\mathbf{S_{12}}$; y ($0\le y\le x$) are in state $\mathbf{S_{11}}$; and $x-y+1$ are in state $\mathbf{S_{22}}$. We denote the total duration of these three categories of time intervals as U, V, and W ($=t-U-V$), respectively. Then

$$\Pr(G \mid \Theta) = \frac{2}{\theta_2} m_1^{x+1} m_2^{x} \exp[-\frac{2}{\theta_1}V - \frac{2}{\theta_2}W - m_1(U+2V) - m_2(U+2W)] \qquad (2.3)$$

(Beerli and Felsenstein, 1999; Hey and Nielsen, 2007). Swapping θ_1 with θ_2, m_1 with m_2 gives the probability of a genealogy in which the coalescent event happens in Pop1.

The exponential function in (2.3) depends only on five variables: x, y, U, V and W. The total probability of a group of genealogies, which share the same value for these five variables, can be calculated by permutation and convolution.

$$\Pr(x,y,U,V,W \mid \Theta) = \begin{cases} 2^{x+1} \dfrac{x!}{(x-y)!\,y!} \dfrac{U^x}{x!} \dfrac{V^{y-1}}{(y-1)!} \dfrac{W^{x-y}}{(x-y)!} m_1^{x} m_2^{x} f(U,V,W,\Theta), & if\ y \ge 1 (V>0) \\ 2^{x+1} \dfrac{U^x}{x!} \dfrac{W^x}{x!} m_1^{x} m_2^{x} f(U,V,W,\Theta), & if\ y = 0 (V=0) \end{cases}$$

$$where\ f(U,V,W,\Theta) = \frac{2m_1}{\theta_2} f_1(U,V,W,\Theta) + \frac{2m_2}{\theta_1} f_2(U,V,W,\Theta)$$

$$f_1(U,V,W,\Theta) = \exp[-\frac{2}{\theta_1}V - \frac{2}{\theta_2}W - m_1(U+2V) - m_2(U+2W)] \qquad (2.4)$$

$$f_2(U,V,W,\Theta) = \exp[-\frac{2}{\theta_2}V - \frac{2}{\theta_1}W - m_2(U+2V) - m_1(U+2W)]$$

Integrating (2.4) over the five variables (under the constraint $U+V+W=t$) gives:

$$\Pr(G^* \mid \Theta) = \iiint_{U+V+W=t} \sum_{x \geq y \geq 0} \Pr(x, y, U, V, W \mid \Theta) = \iiint_{U+V+W=t} g(U,V,W,\Theta) f(U,V,W,\Theta), \text{ for } t < T$$

$$\text{where } g(U,V,W,\Theta) = \begin{cases} \sqrt{\frac{2m_1 m_2 U}{V}} BesselI(0, \sqrt{8m_1 m_2 UW}) BesselI(1, \sqrt{8m_1 m_2 UV}), & \text{if } V > 0 \\ BesselI(0, \sqrt{8m_1 m_2 UW}), & \text{if } V = 0 \end{cases} \quad (2.5)$$

This is the total probability of all the genealogies that share the same coalescent time t ($<T$). To the best of our knowledge, there is no analytical solution to the integration in (2.5). However the function can be precisely approximated by using numerical integration methods. Note that $W=t-U-V$, meaning that the integration is over two variables instead of three.

If the coalescent event happens after T, then at time point T, both genes are either in the same population ($\mathbf{S_{11}}$ $\mathbf{S_{22}}$) or in different populations ($\mathbf{S_{12}}$). The probabilities of these two scenarios, denoted as $Q_0(T,\Theta)$ and $Q_1(T,\Theta)$ respectively, are:

$$Q_0(T,\Theta) = \iiint_{U+V+W=T} g(U,V,W,\Theta) f'(U,V,W,\Theta)$$

$$\text{where } f'(U,V,W,\Theta) = m_1 f_1(U,V,W,\Theta) + m_2 f_2(U,V,W,\Theta) \quad (2.6)$$

$$Q_1(T,\Theta) = \iiint_{U+V+W=T} h(U,V,W,\Theta) f_1(U,V,W,\Theta), where$$

$$h(U,V,W,\Theta) = \begin{cases} 2m_1 m_2 U \sqrt{\frac{m_1 m_2}{VW}} BesselI(1, \sqrt{8m_1 m_2 UW}) BesselI(1, \sqrt{8m_1 m_2 UV}), if\ V,W>0 \\ \sqrt{\frac{2m_1 m_2 U}{V}} BesselI(1, \sqrt{8m_1 m_2 UV}) \quad , if\ V>0 \& W=0 \\ \sqrt{\frac{2m_1 m_2 U}{W}} BesselI(1, \sqrt{8m_1 m_2 UW}) \quad , if\ V=0 \& W>0 \\ 1 \quad , if\ V=0 \& W=0 \end{cases} \quad (2.7)$$

Both Q_0 and Q_1 can be evaluated by numerical integration. And the probability of all genealogies sharing coalescent time t ($>T$), is:

$$\Pr(G^* \mid \Theta) = [Q_0(T,\Theta) + Q_1(T,\Theta)] \frac{2}{\theta_A} \exp[-\frac{2}{\theta_A}(t-T)] \text{ , for } t>T. \quad (2.8)$$

Note that in both (2.5) and (2.8), swapping parameter θ_1 and m_1 with θ_2 and m_2 does not change the value of the functions. This suggests that the likelihood surface is symmetric and sampling one sequence from each population won't provide enough resolution for estimating population parameters. Also, in the case when both migration rates are close to zero, it is impossible to estimate the size of sampled populations when each locus is sampled once from each population. To permit estimation of all parameters, we consider the case where two genes are sampled from the same population (either Pop1 or Pop2) at additional loci. The probability of these genealogies can be derived and evaluated in essentially the same way as described above (See Supplement Methods).

A computer program was written to implement an adaptive multidimensional integration routine for the two-dimensional integration in (2.5) and (2.7). The adaptive routine estimates a function on a hypercube(s) based on cubature rules, returning an estimate of the integral together with an estimate of the error (Johnson, 2005). After each iteration, the routine picks the hypercube with the largest estimated error and divides it into two. The routine stops after the estimated integral converges. Romberg integration (Press *et al.*, 1992) is then used to integrate over t in (2.2). Both simulated annealing and the downhill simplex method as implemented by PRESS *et al.* (Press *et al.*, 1992) are used to search for the maximum likelihood estimator.

Mutation rate variation

If it is assumed that all loci have the same mutation rate, then none of the variation that is observed among loci is considered to be caused by variation in the mutation process. We implement this model and identify the method as the "single-rate method" in order to compare it to models that allow for variation in the mutation rate. In general we expect that methods allowing for variation in mutation rate will be preferable. Failing to account for such variation is expected to lead to an overestimate of the variance in the coalescent process as compensation for the lack of variance in mutation, and therefore should introduce significant bias to the estimates of ancestral population size and species divergence time. (Yang, 1997)

In an MCMC application of the IM model additional locus-specific mutation scalars are assigned to each locus (Hey and Nielsen, 2004). During the MCMC run, these mutation scalars are allowed to vary, subjecting to the constraint that the product of all scalars equals 1. This approach is effective when multiple sequences are sampled (Burgess and Yang, 2008; Hey and Nielsen, 2007). However, as our method uses only two genes at each locus, there is not enough information to partition the variation among loci into that due to variance in coalescent times and that due to variance in mutation rates. Another method uses the average distance from an outgroup sequence to the sample sequences to calculate a relative mutation rate for each locus (Yang, 2002). A problem for this method is that outgroup-sample genealogy shares part of its branch with sample-sample genealogy. This creates an additional correlation between the outgroup-sample distance and the sample-sample distance and may introduce bias to the parameter estimates. To avoid these issues, we develop two new methods that we identify as 'fixed-rate' and 'all-rate' respectively. Both methods rely on sampling an extra pair of sequences, each from an outgroup population, to provide information on mutation rates. When the two chosen outgroup populations have separated from each other for a long time and haven't exchanged genes after initial separation, the variance in coalescent time of these two outgroup sequences is small compared to the long population splitting time and can be neglected. Also, the genealogy of the two outgroup sequences is independent of the genealogy of the sampled sequences. Under such circumstances the distance between the two outgroup sequences becomes a good indicator of the locus-specific mutation rate.

For the fixed-rate method, the distance between the two outgroup sequences, is used to calculate a fixed mutation scalar for each locus. For example, the mutation scalar of the *i*th locus (μ_i) is estimated as the outgroup distance (d_i) divided by the average outgroup distance along all loci ($\bar{d}$),

$$\mu_i = d_i / \bar{d} \qquad (2.9)$$

These fixed rate scalars are used in the calculation of Pr(X|G) during the search for the maximum likelihood.

Outgroups are also used for the all-rate method, but rather than using them to set fixed mutation rate scalars, the divergence between outgroups for each locus is considered as part of the data. The joint probability of the data is found by assuming a Gamma (or Uniform) prior, P(μ_i), for the scalars and by integrating over them;

$$\Pr(X, X_O \mid G, T_O) = \int \Pr(X \mid G, \mu_i) \Pr(X_O \mid T_O, \mu_i) P(\mu_i) d\mu_i \qquad (2.10)$$

Here the outgroup distance is represented by X_O and the time of common ancestry of the outgroups is T_O. When an Infinite-Sites mutation model is applied (Kimura, 1969), the integration in (2.10) can be solved analytically.

Test on Simulated Data

Simulations assuming a single mutation rate: We tested the performance of the method on two groups of data sets simulated under the six-parameter IM model assuming

a single mutation rate for all loci. The first group were simulated under a model of bi-directional migration, with parameter values θ_1=0.005, θ_2=0.003, θ_A=0.002, m_1=50, m_2=100 and T=0.003. The numerical values for population size and splitting time parameters are much less than one, and the values for the migration rate terms are high, because the mutation rate component of these parameters is assumed to be on a per-basepair scale and thus to be quite low. The second group of data sets was simulated without migration, using population parameter values θ_1=0.005, θ_2=0.003, θ_A=0.002, m_1=0, m_2=0 and T=0.003. These parameter values describe a history in which the divergence time was fairly long ago, relative to population size (i.e. the ratio of the population size parameter to the divergence time parameter is on the order of 1). This means that considerable genetic drift will have occurred following population separation and the large majority of genealogies are expected to coalesce before the splitting time. To examine how much data the method requires we simulate data sets with different numbers of loci. For each data set, two genes are sampled at each locus from one of three source types: type "12", where one gene is sampled from each population; and types "11" and "22", for samples where each gene comes from the same population. We expect loci of the "12" type to provide more information on ancient population history (i.e. θ_A and T) and loci of the "11" and "22" types to provide more information on recent population history (i.e. θ_1 and θ_2). Thus, in addition to varying the total number of sampled loci, we also examine the effect of varying the size of the three categories of samples. In total we simulated data with 9 different sets of category sizes (Table 2.1). For each combination of population parameters and category sizes, we simulated 10 data sets. Locus length is fixed at 1000 basepairs. Data was simulated assuming an infinite-sites mutation model

(i.e. all mutations at different points in the sequence) and without recombination, with each locus having an independent coalescent history (i.e. free recombination between loci).

After simulation, each data set was analyzed by searching for the joint maximum likelihood estimate (MLE) for all six parameters. The precision standard of the numerical integration routines is set at 10^{-6} for the log-likelihood of a single locus. This means that for a data set of 10,000 loci, our calculated log-likelihood of the whole data set has an estimated error less than 0.01. For the ten data sets simulated under the same parameters and sample sizes, we calculated the mean and standard deviation of the MLEs (Table 2.1). We also plot the mean MLEs in Figure 2.3, with error bars for the standard deviation (SD). For data sets containing no "11" type of loci, we omit the plots for θ_1 estimates due to their having a very large variance.

We analyze the quality of parameter estimates ($\hat{\Theta}$) using two statistics: bias ($E(\hat{\Theta}-\Theta)/\Theta$) and mean square error (MSE, $E((\hat{\Theta}-\Theta)^2)/\Theta^2$). Bias is a measure of accuracy, whereas the mean square error reflects both accuracy and precision. Since both statistics are scaled by the true value of the parameter, we omit the calculation when the true value is zero.

Mutation rate variation: For each vector of parameter values, 10 data sets were simulated with mutation rate variation. Each data set consists of 10,000 loci (2500 type "11", 5000 type "12", and 2500 of type "22"). A mutation scalar is assigned to each locus at the start of the simulation. These scalars are generated from a Gamma(15, 15) distribution having a mean value of 1. An extra pair of outgroup sequences is simulated at each locus, using the same mutation scalar. The common ancestor time of the two outgroup sequences (T_O) is set to 0.015, which is 5 times the value for T used in the simulations. Each data set is analyzed using both the fixed-rate and the all-rate methods. For the all-rate method we considered four different prior distributions of mutation rates. First we applied three gamma priors, all with the same mean of 1.0, but with different variances (0.10, 0.67 and 0.05). We also considered a uniform prior ($U(0,\infty)$). This prior is attractive because it is uninformative, however it is an improper prior with an infinite mean.

RESULTS

Accuracy of estimates

The means and standard deviations of parameter estimates are listed in Table 2.1 and shown in Figure 2.3. Biases and MSEs are listed in Table 2.2. With an input of 10,000 loci distributed across all three types of samples (Table 2.1and Figure 2.3), the method generates estimates that are quite close to the true values of the parameters, with all true values falling in the range of one SD away from the mean MLE. For set III simulations the mean MLEs for migration rates have a bias <14% and the mean MLEs for the other parameters all have a bias <2% (Table 2.2). As expected, the quality of estimates goes down with decreasing total number of loci. For data sets of 1000 loci (set II), the true parameter values still fall within one SD of the mean, but with considerably larger MSEs. Similarly for data sets of only 100 loci (set I) the MSEs are much larger, although bias for most parameters is still low. In this case the mean MLEs for m_1 and θ_A have an estimated bias of 60.7% and 14.3% respectively.

In addition to the effect of the total number of loci, we see that the quality of parameter estimates depends on the numbers of the three categories of loci. As expected the estimates of θ_A and T are strongly affected by the number of type "12" loci. When this is set to 5000, the method provides quite accurate estimates for θ_A and T (all biases<2.5% and all MSE<0.005), even when the data set contains no type "11". We also see that

estimates of θ_1 and θ_2 improve quickly with more "11" and "22" type of loci, respectively and that accurate estimation of m_1 and m_2 requires high numbers of all three types of loci.

To examine more closely the way that likelihood varies with each parameter we estimated the profile likelihood function and 95% confidence intervals for each parameter for two randomly picked 10,000 locus data sets. The profile likelihood is the maximized likelihood function conditioned on a selected focal parameter of interest. Figure 2.4 shows the profile likelihood curves, and Table 2.3 shows the 95% confidence intervals (CI) calculated from these curves based on the standard assumptions of a likelihood ratio test. The result shows that, for both parameter sets, all true parameters fall in range of the 95% CI. And for data sets simulated with positive migration, we can reject the hypothesis of no migration based on the fact that 0 falls outside of the 95% CI for both m_1 and m_2. These curves also reveal some issues that arise for models that include migration. In the first place the 95% CIs for $\theta_1, \theta_2, \theta_A$ and T are narrower for data sets simulated without migration. Secondly, the confidence intervals for m_1 and m_2 are relatively wider than for the other parameters.

Mutation Scalar Methods

In the simulation studies described above, all data were simulated with a single mutation rate for all loci. However for real data the substitution rates vary across the chromosome, and neglecting such variance may result in misleading estimates. To address this we analyzed data simulated under a model in which mutation rates were

sampled from a Gamma(15,15) distribution and then analyzed these data under both the fixed-rate method and the all-rate method.

Results shown in Table 2.4 and Figure 2.5 confirm that neglecting the variance in mutation rates can lead to poor estimation. The single-rate method tends to overestimate migration rates and ancestral population size while underestimating population splitting time and population sizes for sampled populations. As Table 2.5 shows, estimates generated by the single-rate method have the largest bias. The fixed-rate method generally gives better estimates (smaller bias for all parameters except θ_A) than the single-rate method. However, the fixed-rate method still overestimates ancestral population size and underestimates population splitting time. The all-rate method leads to the most accurate estimates, and it appears that using Gamma prior results in slightly better result than using the improper uniform prior. It also appears that using different shape/scalar parameters for the gamma prior has only a small effect on the estimation, as all three gamma priors that were considered lead to similar estimates. Based on these results the all-rate method is the method of choice.

We also looked at profile likelihoods for a small sample of data sets simulated with mutation rate variation and analyzed using the all-rate method with a Gamma (15, 15) prior. Figure 2.6 shows the profile likelihood curves, and from these the 95% confidences intervals were calculated as described above. As for the single-rate results (Table 2.3 and Figure 2.4), all true parameter values fall in the range of 95% CI. For the data set

simulated with non-zero migration, we can reject the hypothesis of no migration because 0 falls out of the 95% CI for both m_1 and m_2. The profile likelihood for the shape/scale parameter of the gamma prior is shown in Figure 2.7, and we note that the estimated MLE for the gamma prior is close to the true value of 15 used in the simulations.

DISCUSSION

In this paper we describe a new likelihood-based inference method for the 'isolation with migration' model. This method resembles Yang's method (Yang, 2002) by using numerical integration to evaluate the likelihood function. Consequently, both methods share the same limitation in that they can't handle data with large samples at each locus. However they can handle data from many independently segregating loci. The situation raises the question of just how sampling should ideally be proportioned: more loci with few gene copies each? Or fewer loci with more gene copies per locus?

In 2006, Felsenstein studied the accuracy of maximum likelihood estimates of effective population size for a single population (Felsenstein, 2006). Using simulated data he found the accuracies of the MLE were well predicted by the formula developed by Fu and Li (Fu and Li, 1993). According to the formula, the accuracy of the estimation is proportional to the number of loci, and approximately proportional to the logarithm of the number of sampled genes at each locus. This result agrees with the conclusion of Pluzhnikov and Donnelly(Pluzhnikov and Donnelly, 1996) that it is optimal to take small samples from populations. Felsenstein noted that this is because the increase in total branch length by sampling extra sequences goes down as the sample size becomes bigger. Although Felsenstein's study was performed on only a single population, the reasoning can be extended to the case of 'isolation with migration' model, especially when the population splitting time is long compared to both extant population sizes and migration rates are not high. Under this scenario, two samples from the same population will most

likely coalesce before they either enter the ancestral population or migrate into the other population. Thus, estimating the effective population size of a sampled population is similar to the case of estimating the size of a single population, favoring samples with a large number of loci with two gene copies from the population for which the size parameter is to be estimated.

The estimates for ancestral population size and population splitting time depend primarily on samples from different populations, in which case they coalesce only after they enter the ancestral population or after one of them migrate into the other population. This process may take a long time unless migration is high. When additional samples are collected, they tend to coalesce with genes from the same population and contribute little to the length of the genealogy. Since the estimation of ancestral population size and population splitting time relies on old coalescent history, it appears that it's better to have a large number of loci with one sample from each population.

Estimating migration rates also benefits from more loci with one sample from each population, because longer branches are more likely to carry migration events. However, the occurrences of migration events can be detected only if two samples from different populations coalesce before entering the ancestral population. This means that it is not possible to estimate migration rates well without also estimating population sizes well. Therefore, in order to achieve good estimates for migration rates, it is preferable to have multiple loci of all three types. This is in agreement with our results from simulation data.

Many statistical methods in population genetics must somehow deal with the question of how much of the variation that is observed among loci is due to variation in the actual mutation rate. Some methods assume that there is no variation of this type and that all base positions have the same mutation rate (Becquet and Przeworski, 2007; Innan and Watanabe, 2006). For data that match this assumption our method, using the single-rate assumption, performs quite well (Figure 2.3). The problem however, for such methods that assume no mutation rate variation, is that when they are applied to real data that do vary in mutation rate, the additional variance will be attributed to variance in the coalescent process.

If indeed loci really do have different mutation rates, then the question arises how best to use information from outgroup species to account for this variation. If outgroup populations have been separated for enough time, the variance in the coalescent process may be ignored and the expected outgroup divergence is proportional to the local mutation. Thus one direct way to estimate a relative mutation scalar for a locus is to use the observed outgroup divergence at that locus. However even if we assume that there is no variation due to the coalescent in the outgroup divergence, the variance of outgroup divergence still includes both a variance among mutation rates and a stochastic variance of the mutation process. By using a fixed mutation scalar derived from the outgroup divergence we are in effect treating all of the variance in outgroup divergence as being due to mutation rate variation. For the purpose of illustration, assume there is actually no

variation in mutation rate among loci. Under these circumstances some loci will still have larger (or smaller) outgroup divergences due to random variation in the mutation process, and this variation will be interpreted as variation in mutation rates. When these variable fixed rates, that were actually sampled from a process with no mutation rate variation, are applied to the model, the introduced variation leads to additional variation in the coalescent times. We identify this effect of overestimating the variance in sample coalescent time as "over-compensation". As a result of over-compensation we expect to overestimate ancestral population size and to underestimate population splitting time. These biases are in agreement with our results on simulated data (Figure 2.4 and Table 2.5). Burgess and Yang also found similar trends in their study (Burgess and Yang, 2008).

An alternative to using a fixed mutation rate for each locus, based on outgroup divergence, is to treat the mutation scalar as a random variable. In this method we consider outgroup divergence as part of the data, and the joint likelihood of sample divergence and outgroup divergence is integrated over a prior distribution for the mutation scalar. This is equivalent to integrating the likelihood function over the posterior distribution of the mutation scalar that is derived from outgroup divergence. Our analyses with this all-rate method, and a prior gamma distribution, yielded estimates with the least bias, compared to results for other ways of handling the mutation rate scalars. We also observed little sensitivity of estimates to the choice of the gamma distribution parameter.

The method described here is designed for data sets with very small samples for very large numbers of loci. Specifically it can be applied in cases where data is available from two genomes, from each of two closely related species. As DNA sequencing techniques advance, we can anticipate growing availability of multiple whole-genome sequences for pairs of recently diverged species. Although we do not yet have two genome sequences from each of two closely related species, we can anticipate some of the issues that will arise when preparing the data for analysis. One large issue is the choice of outgroup species, which need to have been separated for a relatively long time, so that the ancestral polymorphism is small compared to divergence. On the other hand, these populations should not be too far away from the populations under study to guard against the possibility of not sharing in actual mutation rates.

Two other key issues are recombination and selection. Our method follows basic coalescent theory by assuming mutational neutrality, no recombination within loci and free recombination between loci. Violation of these assumptions will impair the validity of the analysis and bring bias to the estimation (Takahata *et al.*, 1995). Thus, loci that undergo selection or recombination during the divergence process needs to be removed from the input data. When multiple (≥2) genome sequences from each population are available, several statistic tests are available for screening for possible selection events, either by comparing nonsynonymous and synonymous substitutions (dN/dS test (Li *et al.*, 1985)), or by comparing polymorphism and divergence (HKA test(Hudson *et al.*, 1987)). To guard against within-locus recombination, we suggest using short sequences. We also

suggest that sequences be taken from genome locations that are separated by sufficient distance so that their evolutionary histories are effectively independent of each other.

TABLES

TABLE 2.1

Mean and standard deviation of maximum likelihood estimates

	# of loci			Mean and Stand Deviation of MLEs					
	"11"	"12"	"22"	θ_1	θ_2	θ_A	m_1	m_2	T
I	25	50	25	0.00429(0.00119)	0.00301(0.00107)	0.00169(0.00107)	80.360(96.111)	100.668(142.789)	0.00303(0.00067)
				0.00556(0.00325)	0.00326(0.00123)	0.00167(0.00087)	3.485(10.092)	19.537(36.135)	0.00323(0.00047)
II	250	500	250	0.00556(0.00074)	0.00297(0.00045)	0.00205(0.00024)	43.927(21.427)	114.992(46.319)	0.00293(0.00018)
				0.00491(0.0004)	0.00300(0.00022)	0.00191(0.00019)	2.267(4.689)	6.393(16.118)	0.00310(0.00012)
III	2500	5000	2500	0.00509(0.00016)	0.00297(0.00015)	0.00200(0.00004)	43.111(6.005)	110.283(16.236)	0.00300(0.00003)
				0.00497(0.00019)	0.00295(0.00010)	0.00202(0.00006)	0.257(0.458)	1.384(2.598)	0.00299(0.00003)
IV	25	5000	25	0.00516(0.00249)	0.00372(0.00151)	0.00202(0.00007)	108.132(56.601)	45.133(88.054)	0.00299(0.00005)
				0.00481(0.00104)	0.00304(0.00031)	0.00200(0.00004)	0.966(1.953)	0.469(1.351)	0.00302(0.00003)
V	250	5000	250	0.00488(0.00069)	0.00299(0.00052)	0.00200(0.00006)	39.792(40.033)	112.329(57.084)	0.00299(0.00004)
				0.00504(0.00045)	0.00291(0.00016)	0.00196(0.00006)	1.005(1.494)	1.284(3.301)	0.00302(0.00003)
VI	25	5000	2500	0.00468(0.00111)	0.00301(0.00015)	0.00196(0.00011)	40.232(16.520)	98.390(17.358)	0.00302(0.00006)
				0.00561(0.0017)	0.00297(0.00006)	0.00200(0.00004)	0.209(0.579)	0.230(0.719)	0.00301(0.00002)
VII	250	5000	2500	0.00504(0.00053)	0.00304(0.00024)	0.00205(0.00007)	49.613(16.590)	100.807(22.905)	0.00297(0.00004)
				0.00499(0.00042)	0.00301(0.00005)	0.00196(0.00009)	0.224(0.490)	0.466(1.04)	0.00302(0.00004)
VIII	0	5000	250	0.03352(0.08029)	0.00297(0.00077)	0.00201(0.00003)	62.361(57.418)	131.58(97.519)	0.00297(0.00005)
				0.15932(0.17869)	0.00299(0.00025)	0.00200(0.00007)	0.736(1.517)	6.212(10.390)	0.00301(0.00005)
IX	0	5000	2500	0.08354(0.15151)	0.00303(0.00016)	0.00200(0.00006)	74.678(46.795)	100.491(18.893)	0.00299(0.00006)
				0.10070(0.18942)	0.00297(0.00007)	0.00198(0.00006)	1.397(2.357)	3.515(5.334)	0.00303(0.00003)
True Parameters				0.00500	0.00300	0.00200	50.000	100.000	0.00300
				0.00500	0.00300	0.00200	0.000	0.000	0.00300

Numbers outside the parentheses are mean MLEs and numbers inside are standard deviations. Upper part of each cell shows the result from data sets simulated with non-zero migration. Lower part shows the result from data sets simulated without migration.

TABLE 2.2

Mean square error and bias of maximum likelihood estimates

	# of loci			MSE and Bias					
	"11"	"12"	"22"	θ_1	θ_2	θ_A	m_1	m_2	T
I	25	50	25	0.0768(-0.143) 0.4349(0.111)	0.1269(0.003) 0.1768(0.086)	0.3123(-0.155) 0.2158(-0.166)	4.0636(0.607) -	2.0389(0.007) -	0.0500(0.011) 0.0304(0.078)
II	250	500	250	0.0343(0.111) 0.0068(-0.018)	0.0227(-0.009) 0.0054(-0.001)	0.0154(0.025) 0.0108(-0.046)	0.1984(-0.121) -	0.2370(0.150) -	0.0041(-0.023) 0.0027(0.032)
III	2500	5000	2500	0.0013(0.018) 0.0015(-0.006)	0.0025(-0.011) 0.0015(-0.018)	0.0003(0.000) 0.0008(0.009)	0.0334(-0.138) -	0.0369(0.103) -	0.0001(-0.001) 0.0001(-0.004)
IV	25	5000	25	0.2487(0.031) 0.0449(-0.039)	0.3118(0.239) 0.0112(0.014)	0.0013(0.008) 0.0005(0.001)	2.6332(1.163) -	1.0764(-0.549) -	0.0003(-0.003) 0.0002(0.006)
V	250	5000	250	0.0194(-0.024) 0.0081(0.008)	0.0304(-0.003) 0.0037(-0.029)	0.0009(0.001) 0.0011(-0.019)	0.6827(-0.204) -	0.3411(0.123) -	0.0002(-0.003) 0.0001(0.006)
VI	25	5000	2500	0.0530(-0.063) 0.1310(0.123)	0.0024(0.004) 0.0005(-0.010)	0.0031(-0.018) 0.0004(-0.001)	0.1473(-0.195) -	0.0304(-0.016) -	0.0005(0.008) 0.0001(0.002)
VII	250	5000	2500	0.0111(0.008) 0.0072(-0.001)	0.0065(0.015) 0.0003(0.003)	0.0018(0.025) 0.0024(-0.021)	0.1102(-0.008) -	0.0525(0.008) -	0.0003(-0.010) 0.0003(0.007)
VIII	0	5000	250	290.36 (5.704) 2229.8 (30.86)	0.0657(-0.009) 0.0070(-0.003)	0.0003(0.007) 0.0013(0.000)	1.3798(0.247) -	1.0507(0.316) -	0.0004(-0.011) 0.0003(0.003)
IX	0	5000	2500	1165.0(15.71) 1801.5(19.14)	0.0029(0.009) 0.0006(-0.012)	0.0010(0.000) 0.0011(-0.012)	1.1195(0.494) -	0.0357(0.005) -	0.0004(-0.003) 0.0002(0.009)
True Parameters				0.00500 0.00500	0.00300 0.00300	0.00200 0.00200	50.000 0.000	100.000 0.000	0.00300 0.00300

Bias and MSE are scaled by the true value of the parameter and its square, respectively. Numbers outside the parentheses are MSE and numbers inside are biases. Upper part of each cell shows the result from data sets simulated with non-zero migration. Lower part shows the result from data sets simulated without migration.

TABLE 2.3

Maximum likelihood estimate and 95% confidence interval

	θ_1	θ_2	θ_A	m_1	m_2	T
True Parameters	0.00500	0.00300	0.00200	50.000	100.000	0.00300
MLE	0.00496	0.00314	0.00202	51.334	83.078	0.00298
95%CI	(0.00455,0.00537)	(0.00287,0.00346)	(0.00188,0.00216)	(25.575,79.317)	(50.690,116.563)	(0.00287,0.00308)
True Parameters	0.00500	0.00300	0.00200	0.000	0.000	0.00300
MLE	0.00508	0.00307	0.00198	0.000	0.000	0.00298
95%CI	(0.00479,0.00539)	(0.00292,0.00323)	(0.00186,0.00211)	(0.000,3.021)	(0.000,4.512)	(0.00292,0.00305)

Top part of the table shows the result from a data set simulated with non-zero migration. Bottom part of the table shows the result from a data set simulated without migration.

TABLE 2.4

Mean and standard deviation of maximum likelihood estimates

Model for Mutation Scalar		Mean and Stand Deviation of MLEs					
		θ_1	θ_2	θ_A	m_1	m_2	T
I	Single Rate	0.00469(0.00026)	0.00273(0.00018)	0.00277(0.00007)	63.669(28.205)	130.679(36.067)	0.00267(0.00006)
		0.00449(0.00013)	0.00281(0.00010)	0.00282(0.00005)	17.384(8.511)	7.097(8.940)	0.00266(0.00005)
II	Fixed Rate	0.00512(0.00023)	0.00292(0.00018)	0.00258(0.00011)	47.877(15.839)	117.302(27.573)	0.00277(0.00008)
		0.00493(0.00012)	0.00297(0.00007)	0.00260(0.00005)	2.793(4.105)	1.775(3.435)	0.00277(0.00005)
III	Prior Uniform(0,∞)	0.00519(0.00023)	0.00291(0.00018)	0.00200(0.00013)	39.830(15.773)	112.995(27.577)	0.00288(0.00008)
		0.00490(0.00011)	0.00293(0.00008)	0.00197(0.00006)	0.845(1.722)	0.568(1.796)	0.00292(0.00004)
IV	Prior Gamma(10,10)	0.00517(0.00023)	0.00295(0.00018)	0.00199(0.00012)	43.978(15.527)	108.617(25.925)	0.00300(0.00008)
		0.00496(0.00011)	0.00297(0.00008)	0.00197(0.00005)	0.884(1.864)	0.813(2.530)	0.00303(0.00004)
V	Prior Gamma(15,15)	0.00511(0.00023)	0.00292(0.00017)	0.00200(0.00012)	45.486(14.766)	109.654(24.792)	0.00299(0.00008)
		0.00492(0.00011)	0.00296(0.00008)	0.00199(0.00005)	1.105(2.299)	0.870(2.697)	0.00302(0.00004)
VI	Prior Gamma(20,20)	0.00506(0.00022)	0.00291(0.00017)	0.00202(0.00011)	47.824(15.430)	110.574(24.857)	0.00298(0.00008)
		0.00490(0.00010)	0.00294(0.00007)	0.00202(0.00005)	1.237(2.645)	1.402(3.496)	0.00300(0.00004)
True Parameters		0.00500	0.00300	0.00200	50.000	100.000	0.00300
		0.00500	0.00300	0.00200	0.000	0.000	0.00300

Bias is scaled by the true value of the parameter and MSE is scaled by square of the true value of the parameter. Data are simulated with mutation rate variation and analyzed using different mutation scalar methods. Numbers outside the parentheses are mean MLEs and numbers inside are standard deviations. Upper part of each cell shows the result from data sets simulated with non-zero migration. Lower part shows the result from data sets simulated without migration.

TABLE 2.5

Mean square error and bias of maximum likelihood estimates

Model for Mutation Scalar		MSE and Bias					
		θ_1	θ_2	θ_A	m_1	m_2	T
I	Single Rate	0.0064(-0.061) 0.0113(-0.103)	0.0118(-0.091) 0.0053(-0.065)	0.1496(0.385) 0.1707(0.412)	0.3930(0.273) -	0.2242(0.307) -	0.0128(-0.111) 0.0132(-0.114)
II	Fixed Rate	0.0027(0.023) 0.0008(-0.015)	0.0045(-0.028) 0.0006(-0.011)	0.0870(0.290) 0.0907(0.300)	0.1021(-0.042) -	0.1060(0.173) -	0.0064(-0.076) 0.0060(-0.075)
III	Prior Uniform(0,∞)	0.0037(0.038) 0.0008(-0.019)	0.0047(-0.031) 0.0012(-0.023)	0.0043(-0.001) 0.0009(-0.013)	0.1409(-0.203) -	0.0929(0.130) -	0.0022(-0.039) 0.0008(-0.026)
IV	Prior Gamma(10,10)	0.0033(0.035) 0.0005(-0.008)	0.0039(-0.016) 0.0007(-0.008)	0.0037(-0.007) 0.0009(-0.014)	0.1109(-0.120) -	0.0746(0.086) -	0.0007(0.000) 0.0003(0.011)
V	Prior Gamma(15,15)	0.0027(0.023) 0.0007(-0.015)	0.0040(-0.027) 0.0008(-0.013)	0.0034(0.000) 0.0007(-0.006)	0.0954(-0.090) -	0.0708(0.097) -	0.0007(-0.003) 0.0002(0.006)
VI	Prior Gamma(20,20)	0.0020(0.012) 0.0009(-0.021)	0.0043(-0.031) 0.0010(-0.019)	0.0032(0.010) 0.0008(0.011)	0.0971(-0.044) -	0.0730(0.106) -	0.0007(-0.006) 0.0002(0.000)
True Parameters		0.00500 0.00500	0.00300 0.00300	0.00200 0.00200	50.000 0.000	100.000 0.000	0.00300 0.00300

Bias and MSE are scaled by the true value of the parameter and its square, respectively. Data are simulated with mutation rate variation and analyzed using different mutation scalar methods. Numbers outside the parentheses are accuracies and numbers inside are biases. Upper part of each cell shows the result from data sets simulated with non-zero migration. Lower part shows the result from data sets simulated without migration.

TABLE 2.6

Maximum likelihood estimate and 95% confidence interval

	θ_1	θ_2	θ_A	m_1	m_2	T
True Parameters	0.00500	0.00300	0.00200	50.000	100.000	0.00300
MLE	0.00520	0.00295	0.00200	52.898	97.784	0.00301
95%CI	(0.00479,0.00565)	(0.00269,0.00324)	(0.00182,0.00215)	(28.592,81.430)	(64.305,131.988)	(0.00290,0.00314)
True Parameters	0.00500	0.00300	0.00200	0.000	0.000	0.00300
MLE	0.00497	0.00300	0.00207	0.000	0.000	0.00294
95%CI	(0.00471,0.00527)	(0.00286,0.00315)	(0.00194,0.00221)	(0.000,3.651)	(0.000,3.551)	(0.00286,0.00331)

Data are simulated with mutation rate variation and analyzed using a all-rate model with a Gamma(15,15) prior. Top part of the table shows the result from a data set simulated with non-zero migration. Bottom part of the table shows the result from a data set simulated without migration. Both data sets are simulated with mutation rate variation.

FIGURE LEGENDS

FIGURE 2.1 'isolation with migration' Model. The demographic parameters are effective population sizes (θ_1, θ_2, and θ_A), gene migration rates (m_1 and m_2) and population splitting time (T).

FIGURE 2.2 Graphic representation of the three possible states for two sampled gene before coalescent. A migration event will result in a switch from one state to another.

FIGURE 2.3 Maximum likelihood estimates of population parameters. Dots in the graph represent the mean maximum likelihood estimates and bars represent the corresponding standard deviations. I-IX stand for the nine combinations of sample sizes as described in Table 2.1. Panel A-C show the result from data sets simulated with non-zero migration. Panel D-F show the result from data sets simulated without migration.

FIGURE 2.4 Profile likelihood curves for population parameters. Panel A-C show the result from a data set simulated with non-zero migration. Panel D-F show the result from a data set simulated without migration.

FIGURE 2.5 Maximum likelihood estimates of population parameters. Data is simulated with mutation rates sampled from a Gamma(15,15) distribution and analyzed using different mutation scalar methods. Dots in the graph represent the mean maximum likelihood estimates and bars represent the corresponding standard deviations. I-VI stand for the six different mutation scalar methods as described in Table 2.4. Panel A-C show the result from data sets simulated with non-zero migration. Panel D-F show the result from data sets simulated without migration.

FIGURE 2.6 Profile likelihood curves for population parameters. Data is simulated with mutation rate variation and analyzed using a all-rate model with a Gamma(15,15) prior. Panel A-C show the result from a data set simulated with non-zero migration. Panel D-F show the result from a data set simulated without migration.

FIGURE 2.7 Profile likelihood curves for gamma parameter of mutation scalar prior. Panel A shows the curve from a data set simulated with non-zero migration. Panel B shows the curve from a data set simulated without migration.

FIGURES

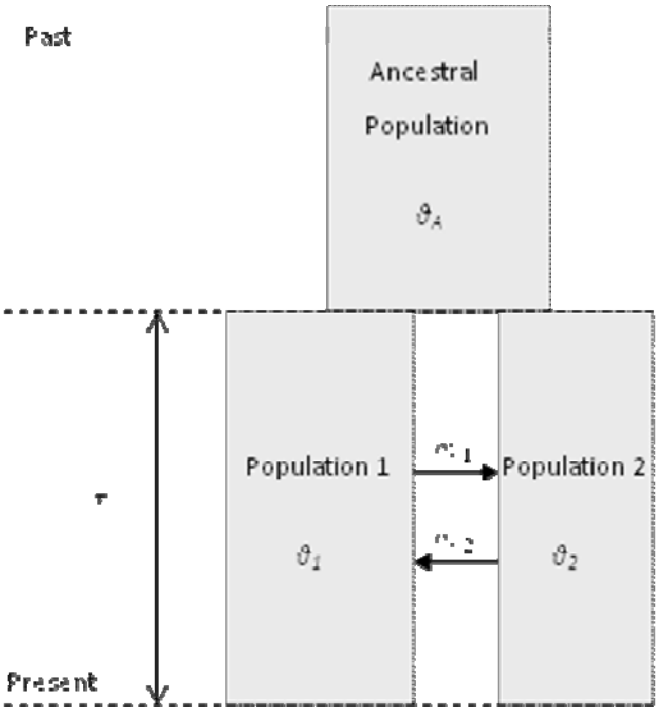

FIGURE 2.1

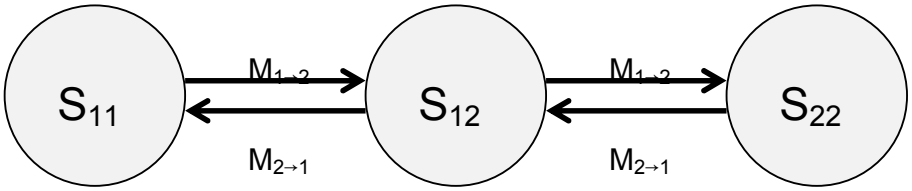

FIGURE 1.2

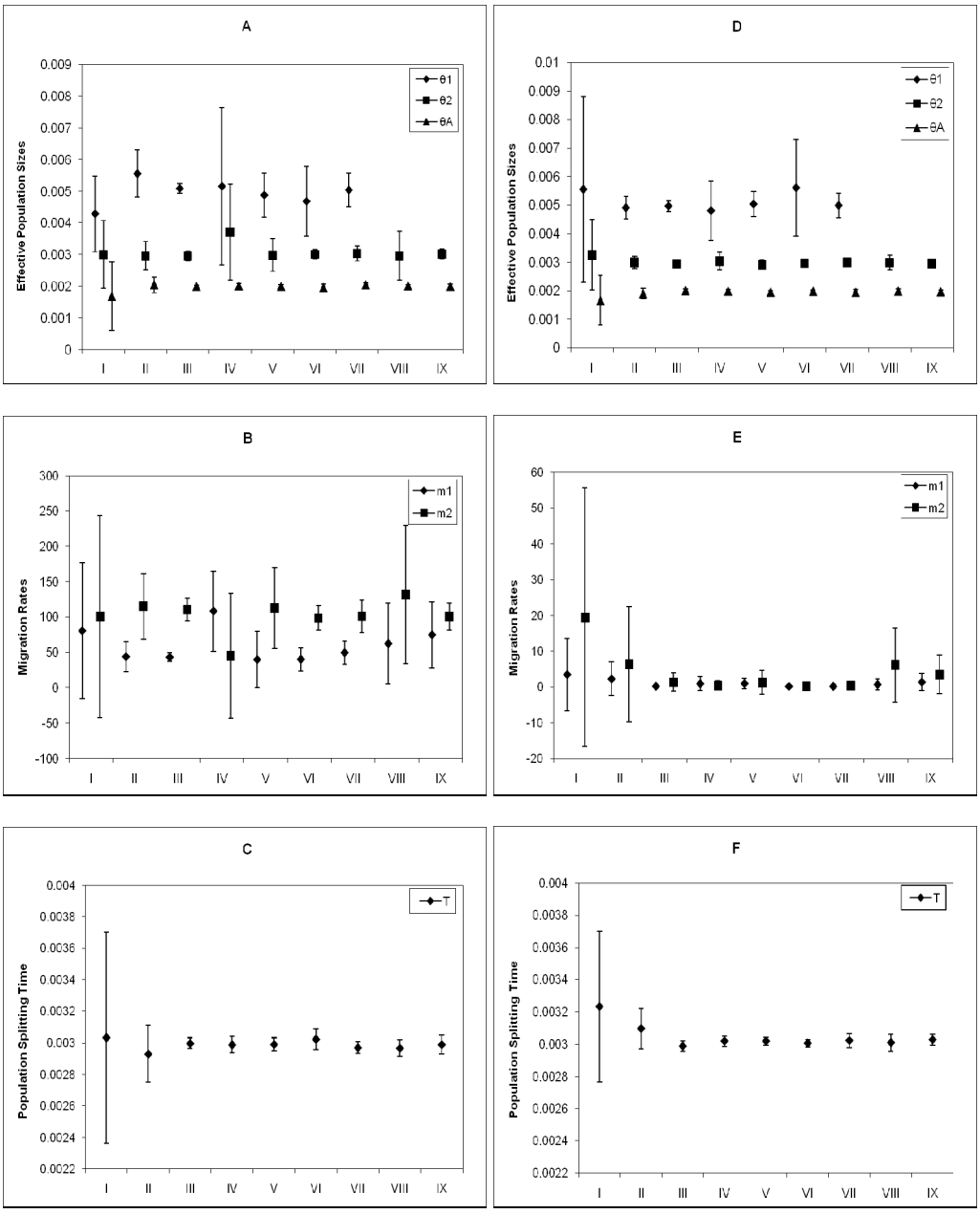
A
Effective Population Sizes
θ1
θ2
θA
D
Effective Population Sizes
θ1
θ2
θA
B
Migration Rates
m1
m2
E
Migration Rates
m1
m2
C
Population Splitting Time
T
F
Population Splitting Time
T
I
II
III
IV
V
VI
VII
VIII
IX

FIGURE 2.3

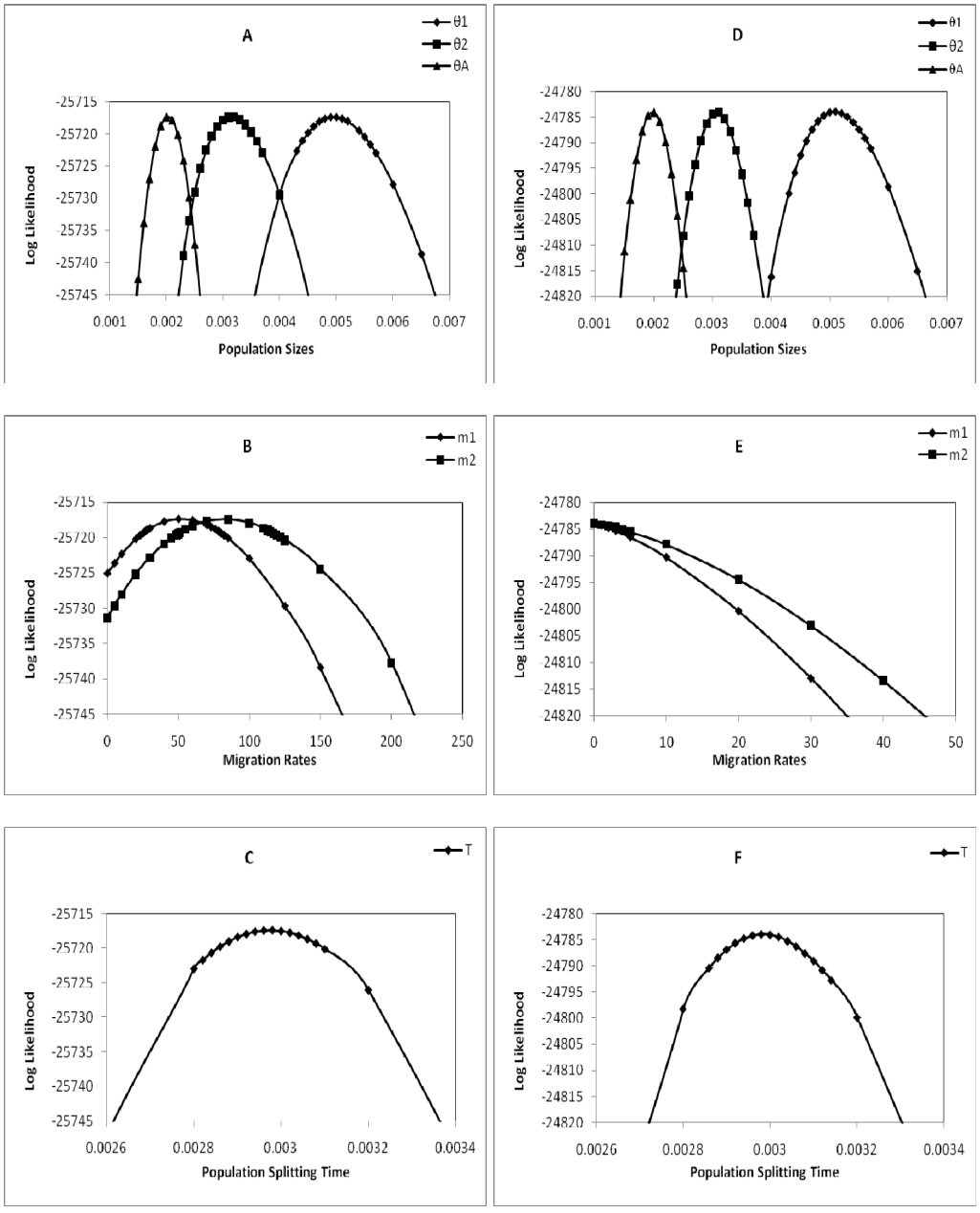
A
θ1
θ2
θA
Log Likelihood
-25715
-25720
-25725
-25730
25735
-25740
-25745
0.001
0.002
0.003
0.004
0.005
0.006
0.007
Population Sizes
D
θ1
θ2
θA
Log Likelihood
-24780
-24785
-24790
-24795
-24800
24805
-24810
-24815
-24820
0.001
0.002
0.003
0.004
0.005
0.006
0.007
Population Sizes
B
m1
m2
Log Likelihood
-25715
-25720
-25725
-25730
-25735
-25740
-25745
0
50
100
150
200
250
Migration Rates
E
m1
m2
Log Likelihood
-24780
-24785
-24790
-24795
-24800
24805
24810
24815
24820
0
10
20
30
40
50
Migration Rates
C
T
Log Likelihood
-25715
-25720
-25725
-25730
-25735
-25740
-25745
0.0026
0.0028
0.003
0.0032
0.0034
Population Splitting Time
F
T
Log Likelihood
-24780
-24785
-24790
-24795
-24800
-24805
-24810
-24815
-24820
0.0026
0.0028
0.003
0.0032
0.0034
Population Splitting Time

FIGURE 2.4

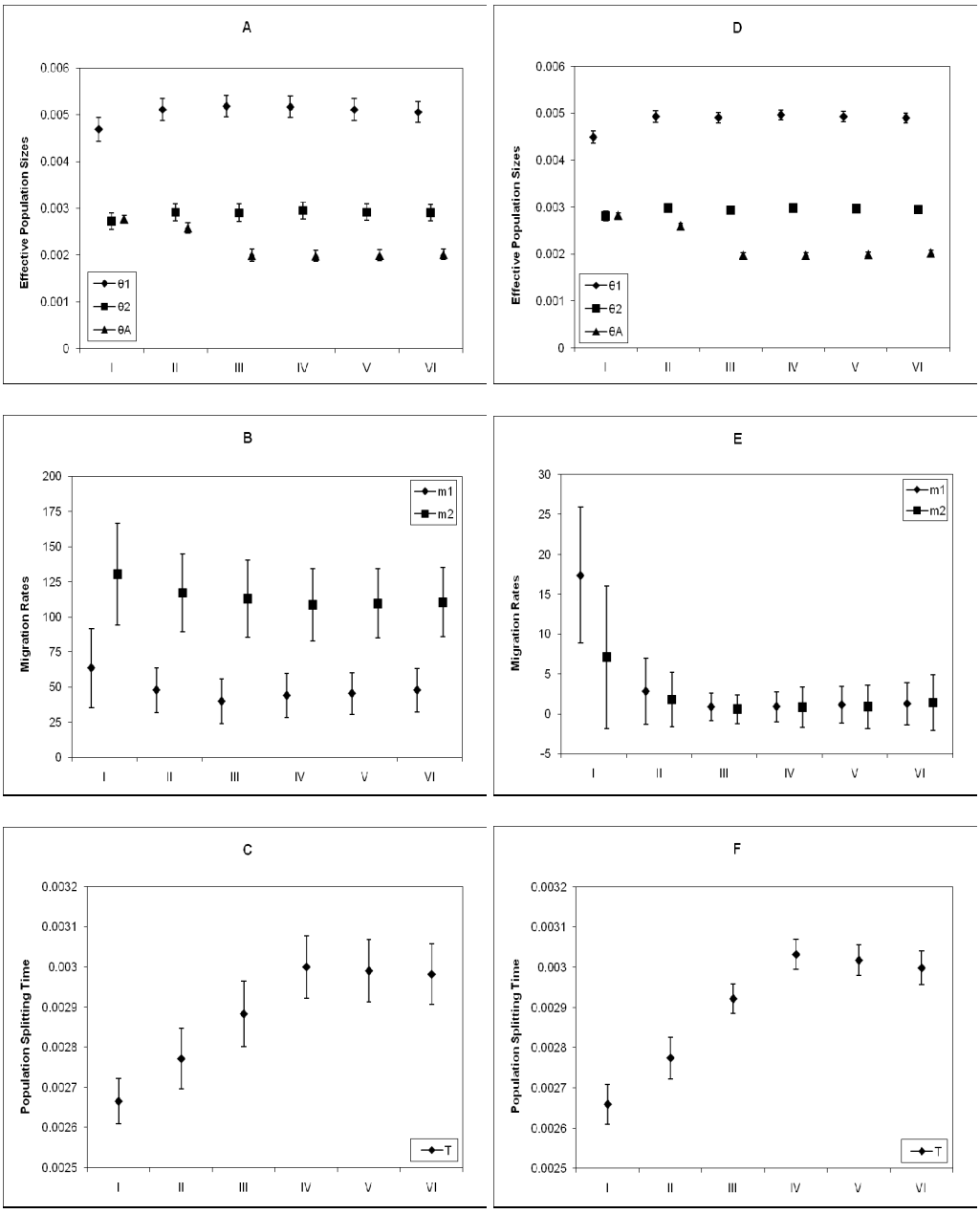
A
Effective Population Sizes
0.006
0.005
0.004
0.003
0.002
0.001
0
θ1
θ2
θA
I
II
III
IV
V
VI
D
Effective Population Sizes
0.006
0.005
0.004
0.003
0.002
0.001
0
θ1
θ2
θA
I
II
III
IV
V
VI
B
Migration Rates
200
175
150
125
100
75
50
25
0
m1
m2
I
II
III
IV
V
VI
E
Migration Rates
30
25
20
15
10
5
0
-5
m1
m2
I
II
III
IV
V
VI
C
Population Splitting Time
0.0032
0.0031
0.003
0.0029
0.0028
0.0027
0.0026
0.0025
T
I
II
III
IV
V
VI
F
Population Splitting Time
0.0032
0.0031
0.003
0.0029
0.0028
0.0027
0.0026
0.0025
T
I
II
III
IV
V
VI

FIGURE 2.5

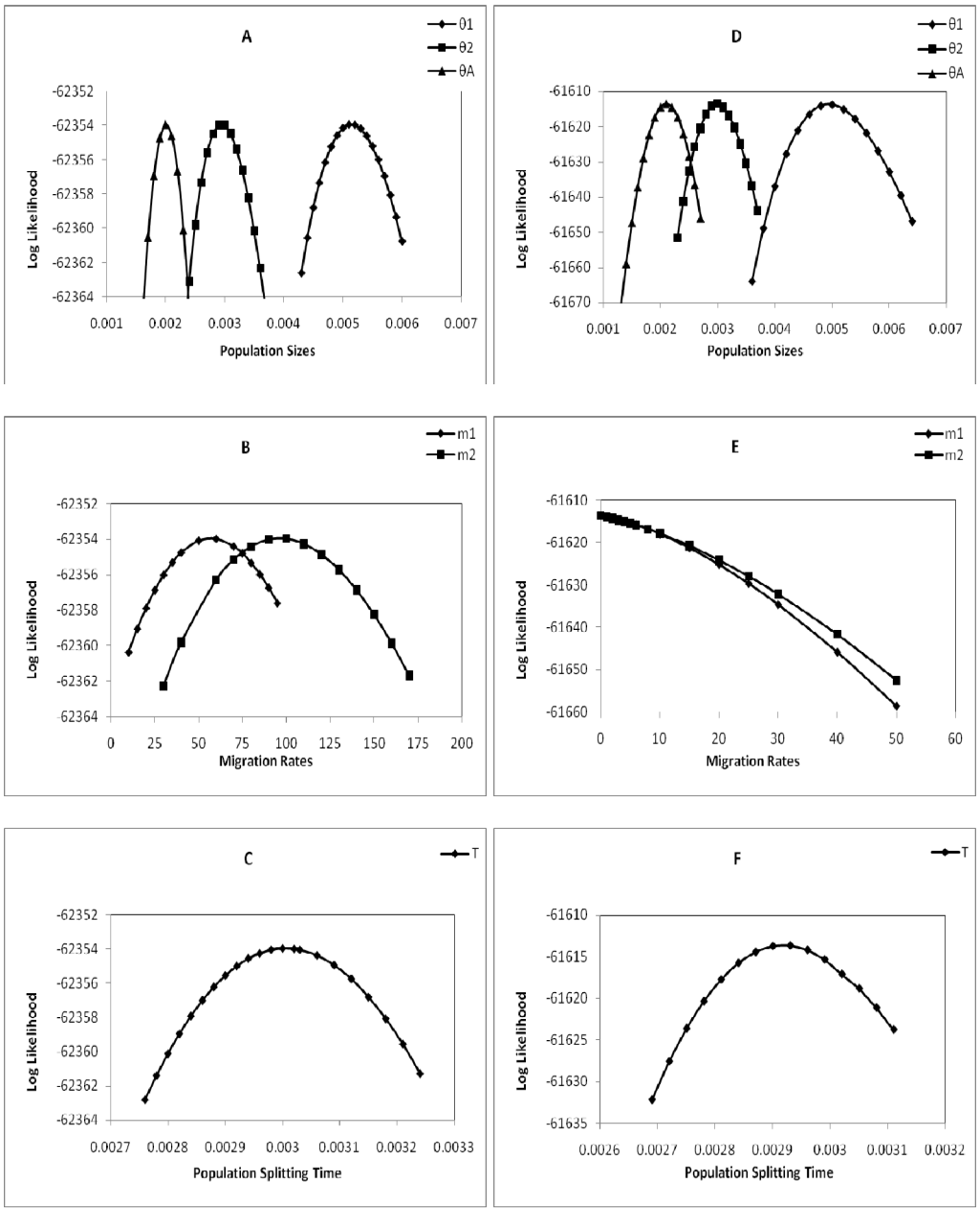
A
θ1
θ2
θA
Log Likelihood
Population Sizes
D
θ1
θ2
θA
Log Likelihood
Population Sizes
B
m1
m2
Log Likelihood
Migration Rates
E
m1
m2
Log Likelihood
Migration Rates
C
T
Log Likelihood
Population Splitting Time
F
T
Log Likelihood
Population Splitting Time

FIGURE 2.6

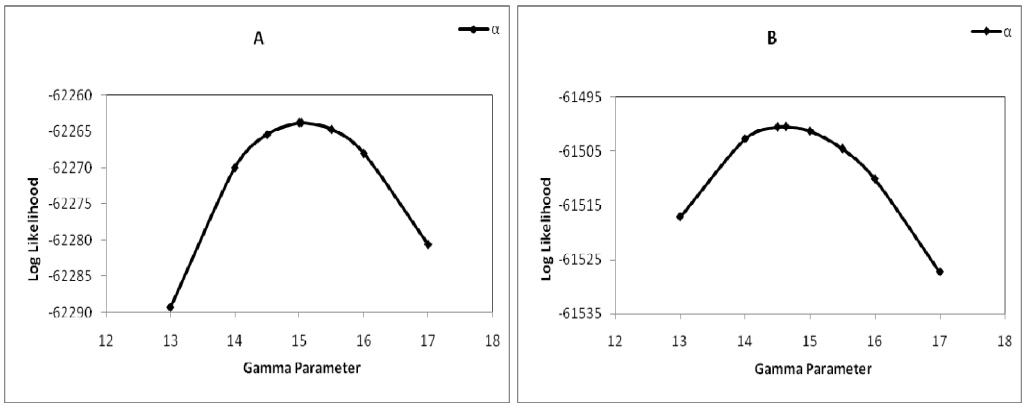
A
α
-62260
-62265
62270
-62275
-62280
-62285
-62290
Log Likelihood
12
13
14
15
16
17
18
Gamma Parameter
B
α
-61495
-61505
-61515
-61525
-61535
Log Likelihood
12
13
14
15
16
17
18
Gamma Parameter

FIGURE 2.7

SUPPLEMENTAL METHODS

Distribution of coalescent time for sample from same population

When two genes are sampled from the sample population, the distribution of coalescent time can be derived in similar way as we showed in the paper. Without losing generality, we assume the both genes are sampled from population1 (i.e. the staring state is $\mathbf{S_{11}}$). The coalescent event can happen either in population 1, or population 2, or the ancestral population. For the first two scenarios, the coalescent time is less than the population splitting time ($t<T$).

Two genes coalesce in population 1: Before the coalescent, there can only be even number ($2x$, $x\geq 0$) of migration events (x of which being $M_{1\to 2}$ and the other x being $M_{2\to 1}$). Of the $2x+1$ time intervals, x are in state $\mathbf{S_{12}}$, $y+1$ ($0\leq y\leq x$) are in state $\mathbf{S_{11}}$ and $x-y$ are in state $\mathbf{S_{22}}$. We denote the total duration of these three categories of time intervals as U, V, and W ($=t-U-V$), respectively. Then

$$\Pr(G \mid \Theta) = \frac{2}{\theta_1} m_1^{\,x} m_2^{\,x} \exp[-\frac{2}{\theta_1} V - \frac{2}{\theta_2} W - m_1(U+2V) - m_2(U+2W)] \quad (2.11)$$

By permutation and convolution, we get

$$\Pr(G^* \mid \Theta) = \iiint_{U+V+W=t} \sum_{x\geq y\geq 0} \Pr(x,y,U,V,W \mid \Theta) = \iiint_{U+V+W=t} \frac{2}{\theta_1} g_1(U,V,W,\Theta) f(U,V,W,\Theta), \; for\, t < T$$

$$where\ g_1(U,V,W,\Theta) =$$

$$\begin{cases} 2m_1 m_2 \sqrt{\frac{V}{W}} BesselI[1, \sqrt{8m_1 m_2 UW}]\, BesselI[1, \sqrt{8m_1 m_2 UV}], if\ U>0, W>0 \\ \sqrt{\frac{2m_1 m_2 V}{U}} BesselI[1, \sqrt{8m_1 m_2 UV}], if\ U>0, W=0 \\ 1, if\ U=0, W=0 \end{cases} \quad (2.12)$$

$$and\ f(U,V,W,\Theta) = \exp[-\frac{2}{\theta_1}V - \frac{2}{\theta_2}W - m_1(U+2V) - m_2(U+2W)]$$

Two genes coalesce in population 2: Before the coalescent, there must be $2x+2$ ($x\geq 0$) migration events ($x+2$ of which being $M_{1\to 2}$ and the other x being $M_{2\to 1}$). Of the $2x+3$ time intervals, $x+1$ are in state $\mathbf{S_{12}}$, $y+1$ ($0\leq y\leq x$) are in state $\mathbf{S_{11}}$ and $x-y+1$ are in state $\mathbf{S_{22}}$. We denote the total duration of these three categories of time intervals as U, V, and W ($=t-U-V$), respectively. Then

$$\Pr(G \mid \Theta) = \frac{2}{\theta_2} m_1^{\,x+2} m_2^{\,x} \exp[-\frac{2}{\theta_1}V - \frac{2}{\theta_2}W - m_1(U+2V) - m_2(U+2W)] \quad (2.13)$$

By permutation and convolution, we get

$$\Pr(G^* \mid \Theta) = \iiint_{U+V+W=t} \sum_{x\geq y\geq 0} \Pr(x,y,U,V,W \mid \Theta) = \iiint_{U+V+W=t} \frac{2}{\theta_2} g_2(U,V,W,\Theta) f(U,V,W,\Theta), \; for\, t < T$$

$$where\, g_2(U,V,W,\Theta) = 2m_1^{\,2} \sqrt{\frac{V}{W}} BesselI[0, \sqrt{8m_1 m_2 UW}]\, BesselI[0, \sqrt{8m_1 m_2 UV}]$$

(2.14)

Two genes coalesce in ancestral population: If the coalescent event happens after *T*, then at time point *T*, both genes are either in the same population ($\mathbf{S_{11}}$ $\mathbf{S_{22}}$) or in different populations ($\mathbf{S_{12}}$). The probabilities of these two scenarios, denoted as $Q_0(T,\Theta)$ and $Q_1(T,\Theta)$ respectively, are:

$$Q_0(T,\Theta) = \iiint_{U+V+W=T} (g_1(U,V,W,\Theta) + g_2(U,V,W,\Theta)) f(U,V,W,\Theta) \quad (2.15)$$

$$Q_1(T,\Theta) = \iiint_{U+V+W=T} h(U,V,W,\Theta) f(U,V,W,\Theta), where$$

$$h(U,V,W,\Theta) = \begin{cases} m_1 \sqrt{\frac{8m_1m_2U}{W}} Bessel[0, \sqrt{8m_1m_2UV}] Bessel[1, \sqrt{8m_1m_2UW}], if\ W > 0 \\ 2m_1 Bessel[0, \sqrt{8m_1m_2UV}], if W = 0 \end{cases}$$

(2.16)

And the probability of all genealogies with coalescent time t ($>T$), is:

$$\Pr(G^* \mid \Theta) = [Q_0(T,\Theta) + Q_1(T,\Theta)] \frac{2}{\theta_A} \exp[-\frac{2}{\theta_A}(t-T)],\ for\ t > T \quad (2.17)$$

Chapter Three

Estimating Population Parameters of Human- Chimpanzee Divergence

ABSTRACT

The divergence of the human and chimpanzee lineages was a pivotal event in human evolutionary history. In this study, we developed a maximum-likelihood (ML) method for joint estimation of six population parameters, based on whole genome data and the 'isolation with migration' model. We applied this method to the alignment of human, chimpanzee and orangutan genome sequences. We estimated that humans and chimpanzees separated approximately 4.3 Myr ago from an ancestral population with a size of ~37000, similar to values obtained by other methods. Surprisingly we detected a clear signal of unidirectional gene flow from chimpanzee to human (0.002 migrations per generation). We showed that this signal is not an artifact created by the recombination or multiple mutation events. To assess the timing of genetic exchange, we extended the model to include a distinct time interval during which gene flow could occur, following population separation. Likelihood ratio test results show that a model with gene flow restricted to the initial period of time following population separation is favored over a model with constant gene flow, from that time until the present.

INTRODUCTION

An oft- debated question in evolutionary research is whether speciation can happen in the presence of gene flow. According to traditional allopatric speciation theory (Dobzhansky, 1936), speciation takes place after a geographic barrier divides an ancestral population into two subpopulations, so that gene flow between them ceases. The two subpopulations then evolve independently and accumulate incompatible mutations. These mutations, in turn, create reproductive isolation and prevent interbreeding if there is subsequent breakdown of the barrier. Alternative theories that do not require the geographic separation have also been proposed. One of those theories is sympatric speciation (Maynard Smith, 1966). In sympatric theory, two populations can diverge from each other while still inhabiting the same geographic region, a scenario built upon the assumption that speciation can happen despite genetic exchange. Although allopatric theory has been widely used in modeling speciation, several recent studies have supported the sympatric model for a variety of speciation events (Llopart *et al.*, 2005) (Niemiller *et al.*, 2008) (Shaw, 2002) (Emelianov *et al.*, 2004) (Turner *et al.*, 2005) (Forbes *et al.*, 2009). These findings suggest that sympatric speciation theory can serve as a useful alternative to traditional allopatric theory.

Of all speciation events, that between humans and our closest relatives, (common chimpanzees) has drawn the most intense interest. As human beings, we are eager to answer the question "How did our ancestors separate from chimpanzees to form a separate species?" Although there is no evidence for recent interbreeding between the

two populations, we cannot rule out the possibility of early gene flow, after the initial separation. But most studies of human-chimpanzee speciation (Takahata *et al.*, 1995) (Yang, 2002) (Rannala and Yang, 2003) (Rannala and Yang, 2003) (Burgess and Yang, 2008) have ignored gene flow subsequent to separation, assuming an instantaneous and complete separation. As a result, these studies attribute the large variance detected in human-chimpanzee divergence to a large ancestral population size. For example, in his classic study (Takahata *et al.*, 1995) Takahata estimated that the human-chimpanzee ancestral population size is about 10 times of current human population size. However, the large variance can also be explained by genetic exchange between human and chimpanzee populations, subsequent to separation. Several studies have recently been conducted to test the hypothesis of sympatric speciation. Osada and Wu (Osada and Wu, 2005) compared the human-chimpanzee divergence of coding sequences and intergenic sequences. Their idea was that if gene flow ceases at the same time, along the whole genome, we would expect no differences in human-chimpanzee divergences between the two types of sequences. However, if effective gene flow stops first at genome regions around some set of 'speciation genes', thought to be important in the divergence of the two taxa, we would expect coding sequences to show larger divergence than intergenic sequences, as coding sequences are more likely to serve as 'speciation genes'. Comparing 345 coding and 143 intergenic sequences, they concluded that the divergence times of coding sequences were different from those of intergenic sequences and rejected the null hypothesis of instantaneous speciation. They suggested that there existed a period of genetic exchange in the human-chimpanzee divergence process. In another study, Patterson and colleagues (Patterson *et al.*, 2006) compared 28 Mb of aligned human, chimpanzee sequences. Their results confirmed the large variance of human-

chimpanzee divergence. More importantly, they found that X chromosome sequences showed much less divergence than those on autosomes. They argued that the difference cannot be explained solely by small population size and low mutation rates for the X chromosome. They proposed a speciation model in which humans and chimpanzees initially separated, but subsequently exchanged genes (mostly X-linked genes) via later hybridization events.

There are criticisms one could level at these results. First, both studies were based only on the summary statistics of the human-chimpanzee divergence and did not use all the information in the data. Second, data used in both studies corresponds to less than 1% of the human genome. Last but not least, they did not estimate the level of gene flow. An initial attempt to evaluate gene flow between humans and chimpanzees was made by Innan and Watanabe (Innan and Watanabe, 2006). They developed a maximum likelihood method to estimate a parameter, α, which represents the level and duration of gene flow. They applied the method to a dataset of 170,000 human-chimpanzee orthologs, each with a length of 100 bps. Their results supported an infinite maximum likelihood estimate of α, tantamount to a model with instantaneous speciation. However, Innan and Watanabe did not evaluate the existence of 'speciation genes', as most loci in their dataset came from non-coding regions.

The data studied by Innan and Watanabe covers ~17 Mb, representing less than 1% of the human genome. The model they used also imposes some restrictions on the speciation process. For example, they assumed symmetric migrations in both

directions and constant mutation rates for each locus. To avoid these problems, we developed an elaborate maximum likelihood method. The method is designed for joint estimation of population sizes, migration rates and population splitting time, using large datasets with two samples for each locus. We applied this method to a dataset extracted from the human-chimpanzee-orangutan genome alignment. This dataset includes ~200,000 loci and covers ~228 Mb, roughly 8% of the human autosomal genome. The greater coverage should provide a more adequate distribution of human-chimpanzee divergence.

MODEL AND METHODS

Model

We based our method on the 'isolation with migration' (IM) model (Figure 3.1). The demographic parameters in this model include three population sizes (θ_1, θ_2 and θ_A), one population splitting time (T), and a pair of migration rates (m_1, m_2). The IM model is a nice tool for examining the speciation process, because it accommodates both sympatric and allopatric speciation. With non-zero migration rates, the IM model can be used to study a sympatric speciation process. When both migration rates are set to zero, the model reduces to allopatric speciation. The fact that two models are nested allows us to compare them via likelihood ratio tests.

A genealogy, G, is a bifurcating tree that represents the evolutionary history of sampled sequences. Given the genealogies at all loci, the probability of the data (X), conditioned on a set of demographic parameters (Θ), can be calculated with coalescent theory and appropriate mutation models. The likelihood of Θ is then found by considering (integrating over) all possible genealogies (Ψ),

$$L(\Theta \mid X) = \Pr(X \mid \Theta) = \int_{\Psi} \Pr(X \mid G) \Pr(G \mid \Theta) dG \qquad (3.1)$$

In most cases, this function cannot be solved by analytical means. Recently, however, we have developed a method (Wang and Hey, submitted) that implements multi-dimensional numerical integration to calculate the likelihood.

Some studies have assumed homogeneous mutation rates among loci (Takahata *et al.*, 1995) (Innan and Watanabe, 2006), but substantial variation in mutation rates has been reported by several analysis (Wolfe and Sharp, 1993). Neglecting such variation will result in overestimating the variance of human-chimpanzee divergence and can lead to questionable inference (Yang, 1997). Other studies account for the variation by scaling the human-chimpanzee divergence against that between humans and orangutans (or/and macaques) (Patterson *et al.*, 2006) (Yang, 2002). As we have pointed out elsewhere (Wang and Hey, 2009), a problem with such an approach is that estimating fixed locus-specific mutation rates still leads to an overestimate of human-chimpanzee divergence variation. In addition, the human-chimpanzee genealogy shares part of its branch length with the human-orangutan genealogy, which entails an additional correlation between the two divergences, which may introduce bias into the parameter estimates. To avoid these problems, we propose a new method that uses the human-orangutan divergence (X_O) as part of the data. The joint probability of the data is found by assuming a Gamma prior, P(μ), for the mutation scalars (locus-specific mutation rates, divided by genome average) and by integrating over them,

$$\Pr(X, X_O \mid G, T_O) = \int \Pr(X \mid G, \mu) \Pr(X_O \mid G, T_O, \mu) P(\mu) d\mu \,. \qquad (3.2)$$

Here, T_O represents common human-orangutan ancestry time. Comparing to the human-orangutan speciation time, T_O has a very small variance and is usually considered as a genomic constant. When the 'infinite sites' mutation model (Kimura, 1969) is used, the integration can be solved analytically.

T_O and the shape parameter of the mutation scale distribution can be estimated from human-orangutan divergence separately. Let L be the sequence length and α be the shape/scale parameter of mutation scales distribution. The number of substitutions between a pair of aligned human and orangutan sequences, N_{HO}, follows a Poisson($2\mu L T_O$) distribution, and the mean and variance of the human-orangutan divergence d_{HO} are

$$E(d_{HO}) = \frac{E(N_{HO})}{L} = \frac{E(E(N_{HO} \mid \mu))}{L} = 2T_O \qquad (3.3)$$

$$Var(d_{HO}) = \frac{Var(N_{HO})}{L^2} = \frac{E(Var(N_{HO} \mid \mu)) + Var(E(N_{HO} \mid \mu))}{L^2} = \frac{2T_O}{L} + \frac{4T_O^{\,2}}{\alpha} \,. \qquad (3.4)$$

Given the mean and variance, T_O and α can be jointly extracted from equation (3.3) and (3.4).

Data Preparation

Alignment genomes for human (assembly hg18), chimpanzee (assembly panTro2) and orangutan (assembly ponAbe2) were retrieved from the multiple alignments of eight vertebrate genomes from USCS Genome Bioinformatics site. 1500-bp long loci were extracted from autosomal regions of the alignment. All loci are separated from their closest neighbor by at least 8000 bps. The human sequence of each locus was blasted against the Celera human genome to obtain a second human sequence (Venter, 2003). The four-gamete test (Hudson and Kaplan, 1985) was performed to search for possible within-locus recombination. If a locus fails the test, only the longest segment that reveals no signal of recombination was retained. Repeated sequences were also removed. A total of 201,432 HHCO (two human, one chimpanzee and one orangutan sample) loci were included. In addition, we acquired 69 pairs of chimpanzee sequences from two earlier studies (Yu *et al.*, 2003) (Fischer *et al.*, 2006). The chimpanzee sequences were blasted against the human-chimpanzee-orangutan genome alignment to search for their human and orangutan orthologs. Of the 69 HCCO loci, 57 passed the four-gamete test and were included in our dataset.

A basic assumption of the method is selective neutrality of data. An early study by The Chimpanzee Sequencing and Analysis Consortium detected six genomic regions with significantly reduced human diversity, relative to human-chimpanzee divergence (2005), suggesting strong selective sweeps in these regions in recent evolutionary history. Loci that fall in these regions were discarded from the dataset.

Loci that overlap with the 3.6-Mb MHC region on chromosome 6 were also removed, because MHC genes are believed to be under intense balancing selection (The MHC sequencing Consortium, 1999).

In the next step, loci with excessive numbers of indels (>0.005 per bp) that distinguished the two human sequences were removed from the dataset. Residues next to indels were also masked to reduce the error introduced by alignment uncertainties. Pairwise distances were calculated for each locus. We assume orangutan as the distant outgroup, so if human-orangutan divergence or chimpanzee-orangutan divergence was less than or equal to other distances (i.e., $\text{Min}(d_{HO}, d_{CO}) \leq \text{Max}(d_{HH}, d_{HC}, d_{CC})$), that locus was removed from the dataset.

Our program requires only three sequences at each locus, including one outgroup sequence. For HHCO loci, with 50% chance, one of the two human sequences was removed at random. Otherwise, the chimpanzee sequence was removed. For HCCO loci, the human sequence was removed. Finally, loci with extremely large divergence (d_{HH}/ d_{CC} >0.03, or d_{HC} >0.07, or d_{HO} >0.1) or small length (L<100) were excluded. In total, the input data consist of 97,999 HHO loci, 98,035 HCO loci and 56 CCO loci.

RESULT

Distribution of Distances

We recalculated the pairwise distances from the final input data (196,090 loci). Distribution curves were plotted in Figure3.1. The average human-orangutan divergence is 0.03033 per site (Table 3.1). Based on equation (3.3), the outgroup ancestry time (T_o) is 0.01516, basically half of human-orangutan divergence. The variance of d_{HO} = 8.829E-5. By solving equation (3.4), the shape parameter of the mutation rate distribution is estimated to be 13.51. This estimation is based on the assumption of a constant human-orangutan ancestry time for all loci, while in reality, the ancestry time varies slightly among loci. So it is possible that we are overestimating the variation in mutation rates, in which case the true shape parameter should be a little larger than our estimate.

The Human-chimpanzee divergence distribution has a mean of 0.01167 per site with a variance of 2.820E-5. This mean is slightly smaller than a distance of 0.0123 from two previous studies (Innan and Watanabe, 2006) (2005). The discrepancy can be explained by different data preparation procedures, as we have removed loci with very large d_{HC}.

Human-human distance has a small average 0.00054, consistent with a small effective size of human population. Average distance between 56 pairs of chimpanzee sequences is very small as well, but we cannot assert a small chimpanzee population size with any confidence, as these chimpanzee sequences were sampled from the

western chimpanzee population and may not be representative of common chimpanzees in general.

Estimation of Human-Chimpanzee Divergence

Maximum likelihood estimates from three separate runs using different mutation rate priors were compared. The results reveal only minor difference (Table 3.2), so our method is robust to the choice of mutation rate prior, and we only use the gamma prior with a shape parameter of 15 for further analysis. In all three cases, our method detects one-way migration from chimpanzee to human (as time moves forward), suggesting the existence of gene flow into the human- lineage, subsequent to separation. To test the significance of the gene flow, we fit data to the 'isolation' model (IM with both $m_1 = 0$ and $m_2 = 0$) and searched for a restricted maximum likelihood estimate. As expected, excluding gene flow leads to larger estimates for the ancestral population size and more recent population splitting time (Table 3.3). Without gene flow, the effective chimpanzee population size depends only on the distance between chimpanzee sequences. This results in a much smaller estimate for chimpanzee population size than that from the model with gene flow. The significance of gene flow is tested by likelihood ratio analysis. Let Λ be the difference between the log likelihoods for two models. Because two parameters (m_1 and m_2) are fixed at boundary values in the reduced model, according to Hey (Hey and Nielsen, 2007), -2Λ is expected to follow a composite χ^2 distribution ($0.25\,\chi_0^2 + 0.5\,\chi_1^2 + 0.25\,\chi_2^2$) under the null hypothesis. In our case $-2\Lambda=1445$, so we reject the null hypothesis ('isolation' model) at a significance level of $p << 0.001$. Interestingly, we

find even the model with one-way migration from human to chimpanzee is significantly better than the 'isolation' model.

To calculate the 95% confidence intervals (CI), we plot the profile likelihood curves for the six population parameters (Figure 3.4). Profile likelihood is the maximized likelihood function, conditioned on a selected focal parameter of interest. The 95% confidence intervals were estimated from these curves, based on the standard assumptions of a likelihood ratio test (Table 3.4). The estimates for human and ancestral population size and population splitting time all have very narrow CIs. Estimates for migration rates have relative wider CIs, but zero falls outside of the 95% CI of m_1, confirming that the gene flow from chimpanzee to human is significant. Estimate for chimpanzee population size has the largest CI, due to the small number of CCO loci.

Population parameters in the IM model are all scaled by the average mutation rate μ ($\theta_1 = 4N_1\mu$, $\theta_2 = 4N_2\mu$, $\theta_A = 4\ N_A\mu$, $m_1 = M_1/\mu$, $m_2 = M_2/\mu$, $T = T'\mu$). To convert the estimates, μ is estimated using human-orangutan speciation time as a calibration value. Several studies have estimated this time to range from 13 Myr to 18 Myr (Glazko and Nei, 2003) (Satta *et al.*, 2004). In our study, we use a human-orangutan speciation time of 15 Myr, together with a 20-year generation time (Gage, 1998). We estimate that the average mutation rate is 1.013E-9 per year and 2.026E-8 per generation. Estimates for population parameters are converted using this value (Table 3.4).

Our estimate for human population size is ~6,200. This result is slightly smaller than the size 10,000 estimated by Takahata (Takahata *et al.*, 1995), (7,500 if they had scaled their result using the same 20-year generation time). We argue that our estimate should be closer to the true value, because we used a much larger data set (97,999 pairs of human sequences) than they did (49 pairs). Despite a small average distance between chimpanzee sequences in our dataset, our analysis resulted in a very large estimate for chimpanzee population size (~35,000), which is close to the estimate of Becquet (Becquet and Przeworski, 2007) and almost double the estimates of Caswell (Caswell *et al.*, 2008) and Hey (Hey, 2009).

The human-chimpanzee speciation time is estimated to be approximately 4.3 Myr. This is in agreement with the 4.6 Myr from Takahata (Takahata *et al.*, 1995). Another recent study (Hobolth *et al.*, 2007) reported a speciation time of 4.1 Myr, but they used an 18-Myr human-orangutan divergence time and a 25-year generation time. Other studies have estimated speciation time of 4 (Burgess and Yang, 2008), 5 (Sarich and Wilson, 1973) (Yang, 2002) and 6 Myr (Glazko and Nei, 2003). Our results suggest that the human-chimpanzee ancestral population had a size of ~37,000, which is larger than the result from two early studies (Yang, 2002) (Rannala and Yang, 2003) but smaller than those from two more recent studies (Hobolth *et al.*, 2007) (Burgess and Yang, 2008). Our estimate for ancestral population size is about six times as large as the estimated human population size. This ratio is between the value of 5 reported by by Wall (Wall, 2003) and that of 10 reported by Takahata (Takahata *et al.*, 1995). Finally, our study estimates that the migration from chimpanzee to human ($2N_1m_1$)

happens at a rate of 0.002 migrations per generation, if we allow continous migration from the time of separation until now. This level of the gene flow, despite being significant, is not strong enough to prevent two populations from diverging (Wright, 1931).

It is hard to compare results from multiple studies that use different calibration points, so we focus on the comparison with a recent study by Burgess and Yang (Burgess and Yang, 2008), who fit a large data set (~7.4 Mb) to the 'isolation' model. Using a 15-Myr human-orangutan divergence time and 1.0E-9 per bp mutation rate, they estimated that humans and chimpanzees separated 4 Myr ago from an ancestral population of size 99,000. Relative to their results, we obtain a slightly deeper estimate for speciation time and much smaller estimate for ancestral population size. The directions of the differences were expected, of course, because part of the variance in human-chimpanzee divergence is explained by gene flow between populations in our study.

Simulation Study

To test the accuracy of our method, we simulated ten data sets using a set of parameters ($\theta_1 = 0.0005$, $\theta_2 = 0.003$, $\theta_A = 0.003$, $m_1 = 10$, $m_2 = 0$, $T = 0.005$, $T_O = 0.015$, $\alpha = 15$) close to our estimates from the human-chimpanzee-orangutan genome alignment. Another ten data sets were then simulated using the same parameters, except without gene flow ($\theta_1 = 0.0005$, $\theta_2 = 0.003$, $\theta_A = 0.003$, $m_1 = 0$, $m_2 = 0$, $T = 0.005$, $T_O = 0.015$, $\alpha = 15$). Each simulated data set contains 20,100 loci, each 1000-bp

long, including 10,000 HHO loci, 10,000 HCO loci and 100 CCO loci. Data was simulated using ‘infinite-site’ mutation model and without recombination.

Maximum likelihood estimates were estimated from these simulated data sets. Their means and standard deviation were calculated (Table 3.5, data group I&VI) and plotted in Figure 3.5. The diamonds represent the means and the error bars represent the corresponding standard deviations (SD). As Figure 3.5 shows, our method generates estimates that are quite close to the parametric values, with all true values falling within one SD of the average MLE.

An important assumption followed by our method is that of no recombination within loci, but with an average locus length over 1000 bps, this assumption is not sound. Although we used the four-gamete test to screen the data, only a part of the recombination events can be detected and removed via that test. In order to assess the impact of recombination on our estimation, we simulated 20 data sets, using the same two sets of parameters, but adding a recombination rate of 1.5E-8 per bp per generation. As shown in Figure 3.5 (data group II&VII), violating the assumption of no recombination within a locus will result in slightly overestimating speciation time and underestimating ancestral population size. However, it has minimal effect on estimating human or chimpanzee population sizes. More importantly, it creates no false-positive signals of gene flow.

Burgess and Yang (Burgess and Yang, 2008) suggested that one can examine the effect of recombination by using a shorter segment of each locus. Their idea was that recombination should have a larger impact on longer sequences. To test this idea, we cut the 20 data sets in half and used the first 500 bps of each locus. Our results (Figure 3.5, data group III&VIII) demonstrate that estimates from half-length data sets do have less bias. We then generated and analyzed a half-length data set from our original data. The results of this analysis are listed in Table 3.2. The new estimates are close to those obtained from the full-length data set, except for chimpanzee population size, which might be a random effect caused by small sample size of CCO loci. Everything considered, we concluded that our estimates for human-chimpanzee divergence are robust to the assumption of within-locus recombination.

Our method uses the 'infinite-site' mutation model which assumes no multiple mutations at a single site. To evaluate the impact of multiple mutations, we simulated 20 data sets with the Jukes-Cantor (JC69) mutation model. Maximum likelihood estimates from these data sets are listed in Table 3.5 (data group IV&IX). No deviation from the true values is detected. Thus we concluded that our estimates are also robust to the choice of mutation models.

Another assumption of our method is the genome-wide constant human-orangutan ancestry time. To examine the consequence of violating this assumption, we simulated 20 data sets with a human-orangutan ancestral population size of 0.003. The results of our analysis (Figure 3.5, data group V&X) demonstrate that neglecting the variance in human-orangutan ancestry time can lead to slight underestimation of

human-chimpanzee splitting time, human population size and chimpanzee population size. However, it has minimal effect on the estimation of migration rates and human-chimpanzee ancestral population size.

Extending the IM model

In 'isolation with migration' model, two populations exchange genes at a constant rate after the initial separation. However, it is reasonable to suppose that gene flow between humans and chimpanzees, given its existence, would only last for a limited period of time, subsequent to isolation. To test the possibility of this scenario, we extended the IM model to a 'two-stage migration' model (Figure 3.2.A). The new model has one more time parameter. In this model, gene flow only exists in the initial time period (T_2) following the initial population split, and ceases thereafter (T_1). The 'two-stage migration' model is nested within both IM and 'isolation' models. When T_1 is set to zero, the 'two-stage migration' model becomes identical to the IM model. And when m_1 and m_2 are set to zero, it reduces to the 'isolation' model.

We fit this 'two-stage migration' model to the human-chimpanzee-orangutan dataset. Our maximum likelihood estimates are listed in Table 3.3, and converted parameters are listed in Table 3.6. Again, the method detects a significant ($-2\Lambda=3052$, $p << 0.001$) and much stronger (0.138 migrations per generation) one-way gene flow from chimpanzee to human. The new estimate for human-chimpanzee ancestral population size is ~20,000. This value is about half of the estimate, based on the IM

model (~37,000), and the new estimate for the human-chimpanzee splitting time, 6.6 Myr (T_1'+ T_2'), is ~50% larger than 4.3 Myr estimated from the IM model.

We compared the likelihood of the two MLEs estimates from 'two-stage migration' and IM models. Since the two models differ by one parameter fixed at boundary a value ($T_1 = 0$), we expect -2Λ to follow a composite χ^2 distribution ($0.5\,\chi^2_0 + 0.5\,\chi^2_1$) (Hey and Nielsen, 2007). With -2Λ = 1607, we rejected the standard model at significance level of $p << 0.001$, suggesting that a two-stage migration scenario provides a better approximation to real human-chimpanzee evolutionary scenario.

In the IM model, the ancestral population size is treated as a constant. However, as pointed out by several studies (Patterson *et al.*, 2006), in one third of the cases, human or chimpanzee sequences are closer to gorilla sequences than to each other. This suggests that gorillas split off shortly before human-chimpanzee speciation. So it might not be appropriate to assume the ancestral population size remains constant over the timescale of human-chimpanzee coalescent. To assess the impact of changing ancestral population size, we developed another model (Figure 3.2.B). In this model, the ancestral population size changed from θ_B to θ_A at time T_A before the population splitting time, T_B. This new model has two more parameters than the standard IM model. When $\theta_A = \theta_B$, these two models become identical to each other.

Maximum likelihood estimates and converted population parameters based on 'changing ancestral size' model are listed in Table 3.3 and Table 3.6. The results shows human-chimpanzee speciation started at 3.7 Myr ago. The human-chimpanzee population size is estimated to be ~70,000, which doubles the size estimated based on IM. This size reduces to ~20,000 at 6.6 Myr ago. Weak but significant ($-2\Lambda = 693$, $p << 0.001$) gene flow from chimpanzee to human is detected as well. The likelihood ratio test suggests 'changing ancestral size' model is also better than the standard IM model ($-2\Lambda=1607$) at the significant level $p << 0.001$.

DISCUSSION

In this study, we estimated the demographic parameters of human-chimpanzee speciation from the human-chimpanzee-orangutan genome alignment. We detected a significant unidirectional gene flow from chimpanzees to humans (as time moves forward). Our finding is in contrast with the result of Innan and Watanabe (Innan and Watanabe, 2006), who studied 17,000 100-bp long human-chimpanzee orthologs and found no signal of gene flow. One of the differences between our study and theirs is that we used longer sequences (average length >1100). Using longer sequences will help determine the human-chimpanzee coalescent time more accurately and improve the quality of the estimates. On the other side, estimation based on longer loci is more likely to be influenced by recombination. Nevertheless, our simulation study demonstrates that the gene flow detected by us is unlikely to be an artifact created by the recombination effect. Nor could it be a false-positive consequence of other effects, including multiple mutations and varying outgroup ancestry time. To make a further comparison of 'isolation' model with 'isolation with migration' model, we simulated two data sets. Data set one was simulated with demographic parameters estimated based on the 'isolation with migration' model, but data set two was simulated with those estimated based on the 'isolation' model. We calculated the distribution of human-chimpanzee divergence from the two simulated data sets and plotted both distributions, along with the divergence distribution from the real data (Figure 3.6). The three distributions are very similar to each other, with the distribution in data set one being a little closer to the real data.

The gene flow we detected is unidirectional from chimpanzee population to human population. A possible explanation for unidirectionality is that genes also were contributed by humans to chimpanzees, but that the recipient population has either remained unsampled to date (or later went extinct). A second explanation is that the signal of gene flow from human to chimpanzee is masked by the effect of recombination. In fact, when half-length data were used, we did detect a marginally significant signal of gene flow in this direction.

We also examined two models extended from standard 'isolation with migration' model to the data. Likelihood ratio tests reveal that each of these models is significantly better than the standard IM, but because these two models are not nested, we were not able to compare them directly with each other. We argue that both models may have represented some aspect of the real history of human-chimpanzee speciation history. However, we find the 'two-stage migration' model gives a relatively large estimate (6.7 Myr) for human-chimpanzee speciation time. This time is in agreement with several fossil finds from the Late Miocene, suggesting that the human- chimpanzee speciation might have happened 7 million years ago (Senut *et al.*, 2001) (Brunet *et al.*, 2002) (Brunet *et al.*, 2005).

TABLES

Table 3.1.

Mean and variance of human-human, human-chimpanzee and human-orangutan distances

Divergence	HH	HC	CC	HO
Number of loci	97999	98035	56	196090
Mean	0.00054	0.01167	0.00049	0.03033
Variance	1.35E-06	2.82E-05	1.13E-06	8.83E-05

Table 3.2.

Maximum likelihood estimates from using full length and half length data

Length	Mutation rate prior	θ_1	θ_2	θ_A	m_1	m_2	T
Full	Gamma(10,10)	0.00050	0.00286	0.00299	6.101	0.000	0.00440
	Gamma(15,15)	0.00050	0.00286	0.00298	5.991	0.000	0.00440
	Gamma(20,20)	0.00050	0.00285	0.00298	5.860	0.000	0.00439
Half	Gamma(10,10)	0.00049	0.00180	0.00310	6.346	1.655	0.00434
	Gamma(15,15)	0.00049	0.00179	0.00310	6.371	1.604	0.00434
	Gamma(20,20)	0.00049	0.00181	0.00310	6.346	1.650	0.00434

Table 3.3.

Testing the speciation models

Model	θ_1	θ_2		θ_A	m_1	m_2	T		Log-likelihood
Basic IM	0.00050	0.00286		0.00298	5.991	0.000	0.00440		-1117868.08
Basic IM, m_1=0	0.00052	0.00049		0.00302	0.000	2.936	0.00436		-1118344.11
Basic IM, m_1=m_2=0	0.00052	0.00050		0.00314	0.000	0.000	0.00426		-1118590.41
	θ_1	θ_2		θ_A	m_1	m_2	T_1	T_2	Log-likelihood
Two-Stage Migration	0.00052	0.00248		0.00166	528.703	0.029	0.00313	0.00357	-1117064.45
Two-Stage Migration, m_1=m_2=0	0.00052	0.00050		0.00314	0.000	0.000	0.00426-T_2	0.00426-T_1	-1118590.41
	θ_1	θ_2	θ_B	θ_A	m_1	m_2	T_B	T_A	Log-likelihood
Changing Ancestral Size	0.00051	0.00379	0.00569	0.00166	3.274	0.000	0.00375	0.00289	-1116734.67
Changing Ancestral Size, m_1=m_2=0	0.00052	0.00049	0.00688	0.00180	0.000	0.000	0.00354	0.00276	-1117081.40

Table 3.4.

MLE and estimated 95% confidence interval and converted population parameters

	θ_1	θ_2	θ_A	m_1	m_2	T
MLE	0.00050	0.00286	0.00298	5.991	0.000	0.00440
Estimated 95% CI	(0.00050, 0.00051)	(0.00237, 0.00345)	(0.00297, 0.00301)	(5.761, 6,332)	(0.000, 0.084)	(0.00439, 0.00440)
	N_1	N_2	N_A	$2N_1M_1$	$2N_2M_2$	T' (Myr)
Converted MLE	6204	35266	36768	0.002	0.000	4.341
Converted 95% CI	(6177, 6255)	(29290, 42510)	(36648, 37086)	-	-	(4.336, 4.438)

Table 3.5.

Mean and standard deviation of maximum likelihood estimates for population parameters from simulated data sets

Data group	θ_1	θ_2	θ_A	m_1	m_2	T
I	0.00050 (0.00001)	0.00304 (0.00014)	0.00295 (0.00007)	9.498 (0.919)	0.499 (0.577)	0.00501 (0.00005)
II	0.00050 (0.00001)	0.00298 (0.00011)	0.00226 (0.00009)	9.850 (1.013)	0.009 (0.012)	0.00536 (0.00005)
III	0.00050 (0.00001)	0.00301 (0.0001)	0.00235 (0.00007)	9.535 (0.797)	0.370 (0.520)	0.00532 (0.00004)
IV	0.00051 (0.00001)	0.00303 (0.00014)	0.00289 (0.00005)	10.070 (0.869)	0.337 (0.443)	0.00508 (0.00002)
V	0.00046 (0.00001)	0.00276 (0.00009)	0.00292 (0.00006)	10.487 (1.516)	0.531 (0.522)	0.00455 (0.00003)
True parameter	0.00500	0.00300	0.00300	10.000	0.000	0.00500
VI	0.00050 (0.00001)	0.00292 (0.00012)	0.00294 (0.00005)	0.026 (0.061)	0.008 (0.023)	0.00503 (0.00002)
VII	0.00050 (0.00001)	0.00308 (0.00012)	0.00223 (0.00004)	0.003 (0.010)	0.101 (0.122)	0.00538 (0.00003)
VIII	0.00050 (0.00001)	0.00307 (0.00010)	0.00233 (0.00008)	0.001 (0.002)	0.310 (0.370)	0.00533 (0.00004)
IX	0.00051 (0.00001)	0.00301 (0.00010)	0.00288 (0.00006)	0.152 (0.120)	0.008 (0.023)	0.00508 (0.00004)
X	0.00047 (0.00001)	0.00278 (0.00008)	0.00293 (0.00007)	0.048 (0.065)	0.038 (0.083)	0.00453 (0.00003)
True parameter	0.00500	0.00300	0.00300	0.000	0.000	0.00500

Each group of data is simulated with a different model. I: Standard 'isolation with migration' (IM); II: IM with recombination, full length; III: IM with recombination, half length; IV: IM with multiple mutation; V: IM with varying outgroup ancestry time; VI: Standard 'isolation' model; VII: Isolation with recombination, full length; VIII: Isolation with recombination, half length. IX: Isolation with multiple mutation; X: Isolation with varying outgroup ancestry time.

Table 3.6.

Converted population parameters of two extended IM models

	N_1	N_2	N_B	N_A	$2N_1M_1$	$2N_2M_2$	T_1'/T_B'(Myr)	T_2'/T_A'(Myr)
Two-Stage Migration	6433	30552	-	20471	0.138	0.000	3.090	3.523
Changing Ancestral Size	6242	46730	70250	20429	0.001	0.000	3.696	2.851

FIGURE LEGENDS

Figure 3.1. Standard 'isolation with migration' model. The demographic parameters are effective population sizes (θ_1, θ_2, and θ_A), gene migration rates (m_1 and m_2) and population splitting time (T). The two arrows indicate the direction of migration as time moves backward.

Figure 3.2. Extended 'isolation with migration' models. A) 'two-stage migration' model: gene flow only lasts for a period of time (T_2) following initial population separation and doesn't happen in the time period (T_1) afterwards. B) 'changing ancestral size' model: Ancestral population size changes from θ_A to θ_B at time T_A , before the initial population separation.

Figure 3.3. Distribution curves of three distances. Solid line represents distribution of distances between pair of human sequences. Dash line represents distribution of distance between human and chimpanzee sequences. And dash-dot line represents distribution of distance between human and orangutan sequences.

Figure 3.4. Profile likelihood curve for population parameters estimated from human-chimpanzee-orangutan genome alignment.

Figure 3.5. Maximum likelihood estimates for population parameters estimated from simulated data sets. Dots in the graph represent the mean maximum likelihood

estimates and bars represent the corresponding standard deviations. I-X stand for the demographic models used to simulate data, as described in Table 3.5.

Figure 3.6. Comparing human-chimpanzee divergence distributions. Line with solid diamond marks represents the divergence distribution in human-chimpanzee-orangutan genome alignment. Line with hollow square marks represents the distribution in a data set simulated with parameters estimated from 'isolation with migration' model ($\theta_1 = 0.0050$, $\theta_2 = 0.00286$, $\theta_A = 0.00298$, $m_1 = 5.991$, $m_2 = 0.000$, $T = 0.00440$, $\alpha = 15$). Line with hollow triangle marks represents the distribution in a data set simulated with parameters estimated from 'isolation' model ($\theta_1 = 0.0052$, $\theta_2 = 0.00050$, $\theta_A = 0.00314$, $m_1 = 0.000$, $m_2 = 0.000$, $T = 0.00426$, $\alpha = 15$).

FIGURES

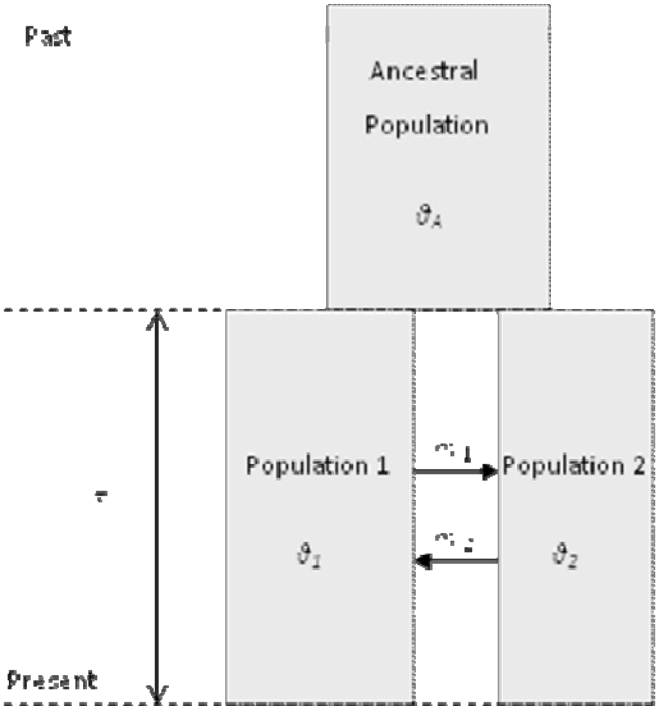

Figure 3.1.

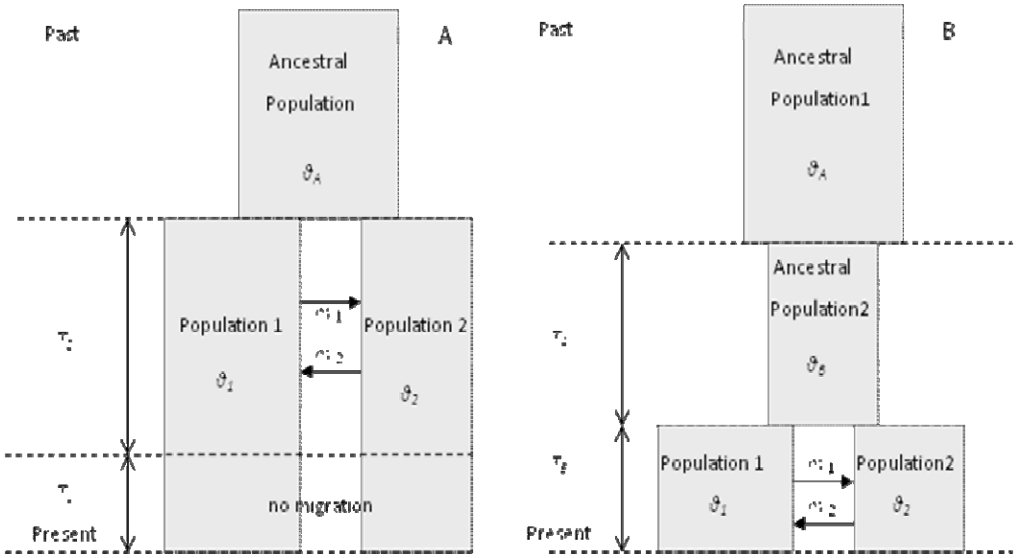
Past
A
Ancestral
Population
Population 1
Population 2
no migration
Present
Past
B
Ancestral
Population1
Ancestral
Population2
Population 1
Population2
Present

Figure 3.2.

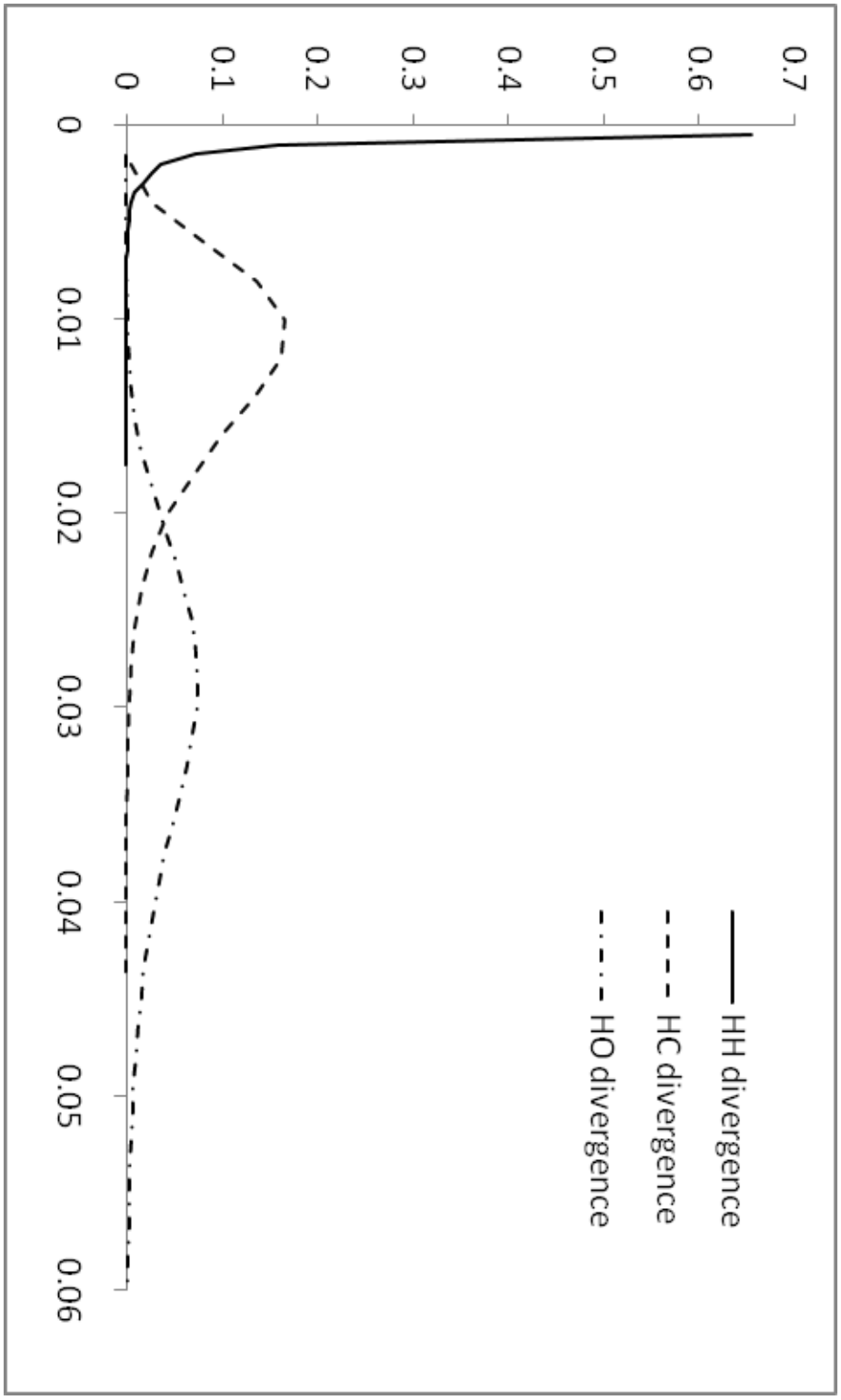
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
0
0.01
0.02
0.03
0.04
0.05
0.06
HH divergence
HC divergence
HO divergence

Figure 3.3

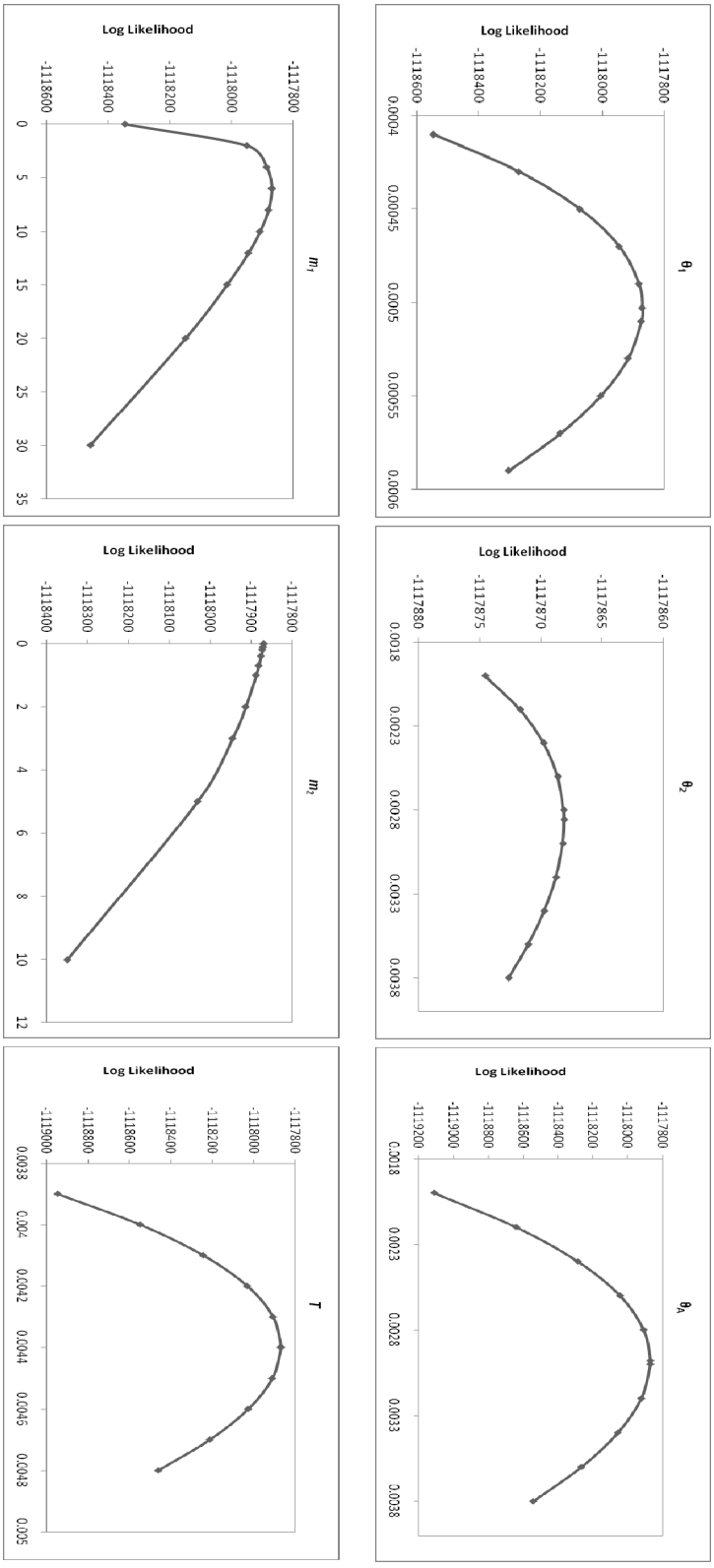

Figure 3.4.

Figure 3.5.

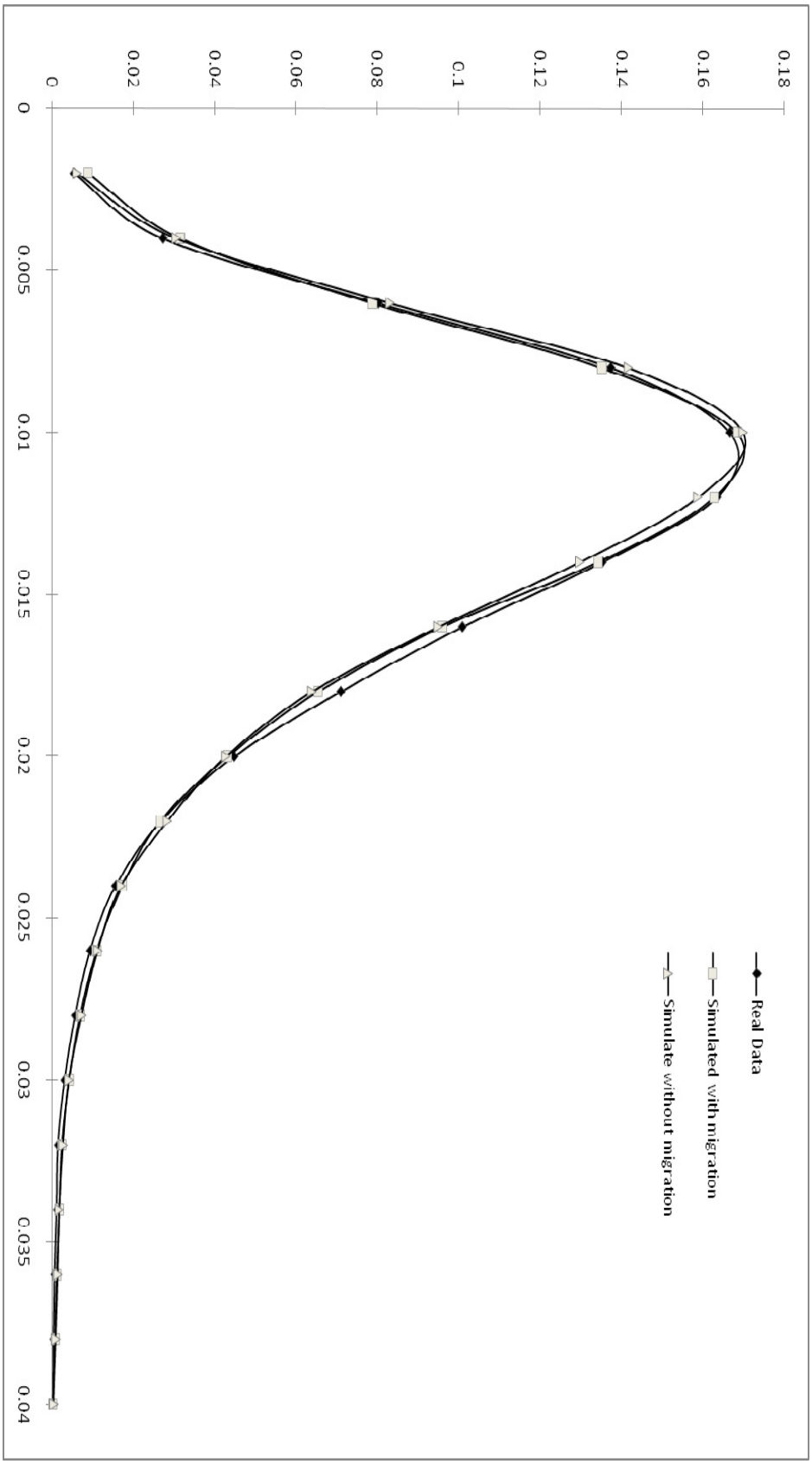
0.18
0.16
0.14
0.12
0.1
0.08
0.06
0.04
0.02
0
0
0.005
0.01
0.015
0.02
0.025
0.03
0.035
0.04
Real Data
Simulated with migration
Simulate without migration

Figure 3.6.

Chapter Four

Conclusions

When a new statistical method is developed to study an evolutionary problem, we must first ask if the method is capable of providing reliable estimates. Extensive tests on simulated datasets demonstrate that our new method generates very accurate estimates from whole-genome alignment data and is a useful tool for estimating the divergence process between closely related species. The simulation results also show that, in order to achieve more accurate estimates, several factors must be taken into account. First, mutation rate variation among loci introduces significant bias into the results, so it is essential to have an outgroup genome. The distance between the sample sequences and the outgroup sequence provides good information about the locus-specific mutation rates. Coupling this information with a newly designed mutation rate variation model greatly improves the accuracy of estimation. Second, it appears that including a number of loci that include a pair of samples from each individual population, helps to estimate the sizes of current populations and the rates of gene flow.

The population parameters of human-chimpanzee divergence were estimated by applying our method to the genome alignment of human, chimpanzee and gorilla. The estimates for the speciation time and population sizes are within the range of several previous studies. A more striking result from this research is the observation of what appears to be a signal of historical gene flow from the chimpanzee to the human population, subsequent to the initial separation. We tested this divergence-with-gene-flow model against the null model of allopatric speciation, and were compellingly able to reject the null model without gene flow. The gene flow estimate for the reverse direction was zero. However, a marginally significant ($-2\Lambda=6.91$, $p<0.005$) signal of gene flow in this direction was discovered, when analyzing half-length data.

The ‘isolation with migration’ model is used in this research. Like many other population genetics models, the IM model is a simplification to real evolutionary history, based on a series of assumptions. Violating these assumptions in practice might lead to questionable results. For this research, a rather important assumption is that of no within-locus recombination. In an early paper (Innan and Watanabe, 2006), Innan and Watanabe reported that recombination could create the signal of gene flow. We argued against their point, because recombination events, by creating independent coalescent histories inside loci, are expect to reduce the variance of human-chimpanzee divergence time among loci. Therefore, recombination should have only attenuated the signal of gene flow. Our simulation confirmed this argument, as analyzing datasets simulated with recombination did not lead to overestimating migration rates. In addition, the effect of violations of some other assumptions, including the ‘infinite sites’ mutation model and constant human-orangutan ancestral time, was studied. In summary, the simulation study found no bias in the estimates for gene flow and only minor deviation in the estimates for other population parameters. As a result, we concluded that the estimate of a non-zero migration rate from chimpanzee to human population is sound.

If gene flow did exist during the human-chimpanzee speciation process, it suggests that divergent selection might have played a role in the evolution and maintenance of reproductive isolation between humans and chimpanzees (Takahasi and Innan, 2008). In recent years, a number of studies (Navarro and Barton, 2003) (Osada and Wu, 2005) (Patterson *et al.*, 2006) have been devoted to search for those

genetic regions subject to the selections that led to the speciation, namely ‘speciation genes’. Although the current research did not attack this problem directly, our new method works as a nice tool in searching for ‘speciation genes’. If a larger estimate of speciation time with two smaller estimates of migration rates is obtained from a group of genes, these genes are more likely to be under the influence of divergent selection. Based on this idea, we are planning to examine the possibility of chromosomal rearrangement regions serving as ‘speciation genes’.

One limitation for the current research comes from the “Trichotomy Problem”. As reported by many studies (Satta *et al.*, 2000) (Chen and Li, 2001) (Rannala and Yang, 2003) (Burgess and Yang, 2008), our next closest relatives, gorillas, split off only shortly before the human-chimpanzee speciation. Given the fact that the divergence times of many human chimpanzee orthologs are even older than the human-gorilla speciation time, it might be better to study the two speciation process in a joint effort. However, without an available gorilla genome, we are currently limited to the human-chimpanzee speciation. With the gorilla genome project in progress, we are planning to extend this method to a three-population model in the future. Some other interesting topics we would like to address are the female-male ratio of mutation rates and female-specific migration rates. Because of the different number of mitotic cycles genes experienced in the two sexes, the mutation rate in males are believed to be much higher than those in female. And in many situations of interspecies hybridization, gene flow is sexually asymmetric, involving males from one species and females from the other species, but not the reverse combination. This leaves a signal of unidirectional gene flow, as we observed in the speciation process for humans and chimpanzees. Both the female-male ratio and the female-specific

migration rates can be evaluated by comparing the divergence between X chromosomal and autosomal rates. A limitation of the current study for attacking this question is that we failed to collect a large enough number of loci from the X chromosome, which (at the moment) consists of a pair of chimpanzee sequences. We are searching for additional X-chromosomal data, and hope to be able to attack this issue in the near future.

Bibliography

Barton NH. 2006. Evolutionary biology: how did the human species form? Curr Biol 16:R647-650.

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

Beerli P, Felsenstein J. 1999. Maximum-likelihood estimation of migration rates and effective population numbers in two populations using a coalescent approach. Genetics 152:763-773.

Beerli P, Felsenstein J. 2001. Maximum likelihood estimation of a migration matrix and effective population sizes in n subpopulations by using a coalescent approach. Proc Natl Acad Sci U S A 98:4563-4568.

Brunet M, Guy F, Pilbeam D, Lieberman DE, Likius A, Mackaye HT, Ponce de Leon MS, Zollikofer CP, Vignaud P. 2005. New material of the earliest hominid from the Upper Miocene of Chad. Nature 434:752-755.

Brunet M, Guy F, Pilbeam D, Mackaye HT, Likius A, Ahounta D, Beauvilain A, Blondel C, Bocherens H, Boisserie JR, De Bonis L, Coppens Y, Dejax J, Denys C, Duringer P, Eisenmann V, Fanone G, Fronty P, Geraads D, Lehmann T, Lihoreau F, Louchart A, Mahamat A, Merceron G, Mouchelin G, Otero O, Pelaez Campomanes P, Ponce De Leon M, Rage JC, Sapanet M, Schuster M, Sudre J, Tassy P, Valentin X, Vignaud P, Viriot L, Zazzo A, Zollikofer C. 2002. A new hominid from the Upper Miocene of Chad, Central Africa. Nature 418:145-151.

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.

Caswell JL, Mallick S, Richter DJ, Neubauer J, Schirmer C, Gnerre S, Reich D. 2008. Analysis of chimpanzee history based on genome sequence alignments. PLoS Genet 4:e1000057.

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.

Dobzhansky T. 1936. Studies on Hybrid Sterility. II. Localization of Sterility Factors in Drosophila Pseudoobscura Hybrids. Genetics 21:113-135.

Emelianov I, Marec F, Mallet J. 2004. Genomic evidence for divergence with gene flow in host races of the larch budmoth. Proc Biol Sci 271:97-105.

Felsenstein J. 1981. Skepticism towards Santa Rosalia, or why are there so few kinds of animals. Evolution 35:124-138.

Felsenstein J. 1988. Phylogenies from molecular sequences: inference and reliability. Annu Rev Genet 22:521-565.

Felsenstein J. 2006. Accuracy of coalescent likelihood estimates: do we need more sites, more sequences, or more loci? Mol Biol Evol 23:691-700.

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

Forbes AA, Powell TH, Stelinski LL, Smith JJ, Feder JL. 2009. Sequential sympatric speciation across trophic levels. Science 323:776-779.

Fu YX, Li WH. 1993. Statistical tests of neutrality of mutations. Genetics 133:693-709.

Futuyma DJ, Mayer GC. 1980. Non-allopatric speciation in animals. Syst. Zool. 29:254-271.
Gage TB. 1998. The comparative demography of primates: with some comments on the evolution of life histories. Annu Rev Anthropol 27:197-221.
Glazko GV, Nei M. 2003. Estimation of divergence times for major lineages of primate species. Mol Biol Evol 20:424-434.
Griffiths RC. 1989. Genealogical-tree probabilities in the infinitely-many-site model. J Math Biol 27:667-680.
Hey J. 2003. Speciation and inversions: chimps and humans. Bioessays 25:825-828.
Hey J. 2006. Recent advances in assessing gene flow between diverging populations and species. Curr Opin Genet Dev 16:592-596.
Hey J. 2009. Isolation with migration models for more than two species: discerning the demographic evolutionary history of chimpanzees. submitted to Mol Biol Evol.
Hey J, Nielsen R. 2004. Multilocus methods for estimating population sizes, migration rates and divergence time, with applications to the divergence of Drosophila pseudoobscura and D. persimilis. Genetics 167:747-760.
Hey J, Nielsen R. 2007. Integration within the Felsenstein equation for improved Markov chain Monte Carlo methods in population genetics. Proc Natl Acad Sci U S A 104:2785-2790.
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, Kaplan NL. 1985. Statistical properties of the number of recombination events in the history of a sample of DNA sequences. Genetics 111:147-164.
Hudson RR, Kreitman M, Aguade M. 1987. A test of neutral molecular evolution based on nucleotide data. Genetics 116:153-159.
Innan H, Watanabe H. 2006. The effect of gene flow on the coalescent time in the human-chimpanzee ancestral population. Mol Biol Evol 23:1040-1047.
Johnson SG. 2005. Adaptint.c, GNU Scientific Library Extensions. In.
Kimura M. 1969. The number of heterozygous nucleotide sites maintained in a finite population due to steady flux of mutations. Genetics 61:893-903.
Kuhner MK, Yamato J, Felsenstein J. 1995. Estimating effective population size and mutation rate from sequence data using Metropolis-Hastings sampling. Genetics 140:1421-1430.
Li WH, Wu CI, Luo CC. 1985. A new method for estimating synonymous and nonsynonymous rates of nucleotide substitution considering the relative likelihood of nucleotide and codon changes. Mol Biol Evol 2:150-174.
Llopart A, Lachaise D, Coyne JA. 2005. Multilocus analysis of introgression between two sympatric sister species of Drosophila: Drosophila yakuba and D. santomea. Genetics 171:197-210.
Lu J, Li WH, Wu CI. 2003. Comment on "Chromosomal speciation and molecular divergence-accelerated evolution in rearranged chromosomes". Science 302:988; author reply 988.
Maynard Smith J. 1966. Sympatric speciation. The American naturalist 100:637-650.
Muller H. 1940. Bearings of the Drosophila work on systematics. In: Huxley JJ, editor. The new systematics. Oxford, UK: Clarendon Press. p 185-268.
Navarro A, Barton NH. 2003. Chromosomal speciation and molecular divergence--accelerated evolution in rearranged chromosomes. Science 300:321-324.

Nei M. 1987. Molecular Evolutionary Genetics. New York: Columbia University Press.
Nielsen R. 2000. Estimation of population parameters and recombination rates from single nucleotide polymorphisms. Genetics 154:931-942.
Nielsen R, Wakeley J. 2001. Distinguishing migration from isolation: a Markov chain Monte Carlo approach. Genetics 158:885-896.
Niemiller ML, Fitzpatrick BM, Miller BT. 2008. Recent divergence with gene flow in Tennessee cave salamanders (Plethodontidae: Gyrinophilus) inferred from gene genealogies. Mol Ecol 17:2258-2275.
Nosil P. 2008. Speciation with gene flow could be common. Mol Ecol 17:2103-2106.
Osada N, Wu CI. 2005. Inferring the mode of speciation from genomic data: a study of the great apes. Genetics 169:259-264.
Patterson N, Richter DJ, Gnerre S, Lander ES, Reich D. 2006. Genetic evidence for complex speciation of humans and chimpanzees. Nature 441:1103-1108.
Pluzhnikov A, Donnelly P. 1996. Optimal sequencing strategies for surveying molecular genetic diversity. Genetics 144:1247-1262.
Press WH, Teukolsky SA, Vetterling WT, Flannery BP. 1992. Numerical Recipes in C - the Art of Scientific Computing, Second ed.
Rannala B, Yang Z. 2003. Bayes estimation of species divergence times and ancestral population sizes using DNA sequences from multiple loci. Genetics 164:1645-1656.
Rice WR, Hostert EF. 1993. Laboratory experiments on speciation : what have we learned in 40 years? Evolution 47:1637-1653.
Rieseberg LH. 2001. Chromosomal rearrangements and speciation. Trends Ecol Evol 16:351-358.
Sarich VM, Wilson AC. 1973. Generation time and genomic evolution in primates. Science 179:1144-1147.
Satta Y, Hickerson M, Watanabe H, O'HUigin C, Klein J. 2004. Ancestral population sizes and species divergence times in the primate lineage on the basis of intron and BAC end sequences. J Mol Evol 59:478-487.
Satta Y, Klein J, Takahata N. 2000. DNA archives and our nearest relative: The trichotomy problem revisited. Molecular Phylogenetics and Evolution 14:259-275.
Senut B, Pickford M, Gommery D, Mein P, Cheboi K, Coppens Y. 2001. First hominid from the Miocene (Lukeino Formation, Kenya). Comptes Rendus De L Academie Des Sciences Serie Ii Fascicule a-Sciences De La Terre Et Des Planetes 332:137-144.
Shaw KL. 2002. Conflict between nuclear and mitochondrial DNA phylogenies of a recent species radiation: what mtDNA reveals and conceals about modes of speciation in Hawaiian crickets. Proc Natl Acad Sci U S A 99:16122-16127.
Takahasi K, Innan H. 2008. Inferring the process of Human-Chimpanzee Speciation. Encycoopedia of Life Sciences April.
Takahata N. 1986. An attempt to estimate the effective size of the ancestral species common to two extant species from which homologous genes are sequenced. Genet Res 48:187-190.
Takahata N, Satta Y, Klein J. 1995. Divergence time and population size in the lineage leading to modern humans. Theor Popul Biol 48:198-221.
The Chimpanzee Sequencing and Analysis Consortium. 2005. Initial sequence of the chimpanzee genome and comparison with the human genome. Nature 437:69-87.

The MHC sequencing Consortium. 1999. Complete sequence and gene map of a human major histocompatibility complex. Nature 401:921-923.
Turner TL, Hahn MW, Nuzhdin SV. 2005. Genomic islands of speciation in Anopheles gambiae. PLoS Biol 3:e285.
Venter JC. 2003. A part of the human genome sequence. Science 299:1183-1184.
Wakeley J. 2008. Complex speciation of humans and chimpanzees. Nature 452:E3-4; discussion E4.
Wall JD. 2003. Estimating ancestral population sizes and divergence times. Genetics 163:395-404.
Wang Y, Hey J. 2009. Estimatig divergence parameters with small samples from a large number of loci. submitted to Genetics.
Wilson IJ, Balding DJ. 1998. Genealogical inference from microsatellite data. Genetics 150:499-510.
Wolfe KH, Sharp PM. 1993. Mammalian gene evolution: nucleotide sequence divergence between mouse and rat. J Mol Evol 37:441-456.
Wright S. 1931. Evolution in Mendelian Populations. Genetics 16:97-159.
Wu CI. 1991. Inference of species phylogeny in relation to segregation of ancient polymorphisms. Genetics 127:429-435.
Yang Z. 1997. On the estimation of ancestral population sizes of modern humans. Genet Res 69:111-116.
Yang Z. 2002. Likelihood and Bayes estimation of ancestral population sizes in hominoids using data from multiple loci. Genetics 162:1811-1823.
Yu N, Jensen-Seaman MI, Chemnick L, Kidd JR, Deinard AS, Ryder O, Kidd KK, Li WH. 2003. Low nucleotide diversity in chimpanzees and bonobos. Genetics 164:1511-1518.

Curriculum Vitae

Yong Wang

Education

B. S. in Biological Science

University of Science and Technology of China, China

1998-2003

B.E. in Computer Science

University of Science and Technology of China, China

1998-2003

M. S. in Statistics

Rutgers University

2003-2008

Ph. D. in Microbiology and Molecular Genetics

Rutgers University

2003-2009

Publications

Wang Y, Hey J. 2009. Estimating population parameters of human-chimpanzee divergence. manuscript in preparation.

Wang Y, Hey J. 2009. Estimating divergence parameters with small samples from a large number of loci. submitted to Genetics.

Won YJ, Wang Y, Sivasundar A, Raincrow J, Hey J. 2006. Nuclear gene variation and molecular dating of the cichlid species flock of Lake Malawi. Mol Biol Evol 23:828-837.

Won YJ, Sivasundar A, Wang Y, Hey J. 2005. On the origin of Lake Malawi cichlids: a population genetic analysis of divergence. Proc Natl Acad Sci USA 102 1:6581-6586.