12-1-2016

Statistical Inference of Genetic Forces Using a Poisson Random Field Model with Non-Constant Population Size

Jianbo Xu
University of Nevada, Las Vegas, jude.xu1989@gmail.com

STATISTICAL INFERENCE OF GENETIC FORCES USING A POISSON RANDOM FIELD MODEL WITH NON-CONSTANT POPULATION SIZE

by

Jianbo Xu

Bachelor of Science
University of Science and Technology of China
2010

A dissertation submitted in partial fulfillment of
the requirements for the

Doctor of Philosophy - Mathematical Sciences

Department of Mathematical Sciences
College of Sciences
The Graduate College

University of Nevada, Las Vegas
December 2016

UNLV | GRADUATE COLLEGE

Dissertation Approval

The Graduate College
The University of Nevada, Las Vegas

November 16, 2016

This dissertation prepared by

Jianbo Xu

entitled

Statistical Inference of Genetic Forces Using a Poisson Random Field Model with Non-Constant Population Size

is approved in partial fulfillment of the requirements for the degree of

Doctor of Philosophy - Mathematical Sciences
Department of Mathematical Sciences

Amei Amei, Ph.D.
Examination Committee Chair

Kathryn Hausbeck Korgan, Ph.D.
Graduate College Interim Dean

Malwane Ananda, Ph.D.
Examination Committee Member

Chih-Hsiang Ho, Ph.D.
Examination Committee Member

Guogen Shan, Ph.D.
Graduate College Faculty Representative

ABSTRACT

STATISTICAL INFERENCE OF GENETIC FORCES USING A POISSON RANDOM FIELD MODEL WITH NON-CONSTANT POPULATION SIZE

by

Jianbo Xu

Dr. Amei Amei, Examination Committee Chair
Associate Professor of Statistics
University of Nevada, Las Vegas, USA

The fidelity of DNA sequence data makes it a perfect platform for quantitatively analyzing and interpreting evolutionary progress. By comparing the information between intraspecific polymorphism with interspecific divergence in two sibling species, the well-established Poisson Random Field theory offers a statistical framework with which various genetic parameters such as natural selection intensity, mutation rate and speciation time can be effectively estimated. A recently developed time-inhomogeneous PRF model has reinforced the original method by removing the assumption of stationary site frequency, but it preserves the condition that the two sibling species share same effective population size with their ancestral species. This dissertation explores a relaxation of this biologically unrealistic assumption by hypothesizing that each of the two descendant species experienced a sudden change in population size at the times of divergence from their most recent common ancestor. Statistical inference of the various genetic parameters are made under a hierarchical Bayesian framework and carried out with a multi-layer Markov chain Monte Carlo sampling scheme. To meet the intensive computational demand, a `R` program is integrated with `C++` code and a parallel executing technique is designed to run the program with multiple CPU cores.

ACKNOWLEDGEMENTS

I would like to express sincerest gratitude to my advisor, Dr. Amei Amei, for her continuous advice, guidance, patience and immense knowledge shared, without which this work would not be possible. She has supported me not only academically but also psychologically through the tough road to finish this dissertation. Thanks to her, I have this priceless opportunity to learn and to grow.

Many thanks to Dr. Chih-Hsiang Ho, Dr. Malwane Ananda, Dr. Kaushik Ghosh, Dr. Hokwon Cho and all faculty members in Department of Mathematical Sciences who helped me during the academic work of past six years. I gained countless benefits from the lectures, seminars and conversations with them.

I would like to thank Dr. Guogen Shan for his kind instruction and inspiration in our collaborative research.

I also thank my friend and senior colleague, Dr. Libo Zhou, for his emotional encouragement during my hardest times.

Finally, I shall thank my families, especially my wife. Her unconditional love and dedication helps me go through the study and life.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

The structure of organisms and their active physiological processes are mainly based on proteins. For most of existing creatures, the genetic information inherited from their ancestors for the synthesis of proteins is contained in a threadlike double-helical molecule named *Deoxyribonucleic acid (DNA)* in the cells (Griffiths et al. (2008)). DNA are polymers (large molecules) consisting of fewer than a hundred to millions or even billions of monomeric units called *nucleotides*. There are four types of nucleotides in DNA, denoted by letters *A, T, C, G* and the two chains of the double helix hold together by complementary pairing of A with T and of G with C. DNA are organized into chromosomes while the region or the segment of chromosomal DNA involved in the cells' production of proteins are called *genes*. The collection of all genes in an organism is called its *genome*. The somatic cells of most plants and animals contain two copies of genome which means their DNA are aligned in paired chromosomes. These creatures are called *diploid*. While the cells of bacteria, algae and most fungi contain only one copy of genome and they are called *haploid*.

The information in genes is used to control the production of proteins according to the following two steps. First, one of the two strands in DNA acts as a template for generating the *messenger ribonucleic acid* (*mRNA*) which is also a sequence of nucleotides. The difference between mRNA and DNA is that the mRNA has nucleotide U instead of T. In other words, DNA determines the sequence of nucleotides in mRNA in such a way that

A in DNA will be paired with U in mRNA. This step is called *transcription*. Second, the nucleotides in mRNA is read in consecutive triplets which correspond to certain amino acids. The resulting string of amino acids, i.e. polypeptides, will then form proteins. Each triplet in mRNA is called a *codon*. Since we have $4^3 = 64$ possible codons but only around 20 amino acids, multiple codons may lead to one amino acid. For instance, AUC, AUU and AUA all encode the amino acid isoleucine. Thus a permanent alternation of nucleotide at a site in DNA, referred as a *mutation*, may result in two cases. The new codon still codes for the same amino acid and alternatively it yields a different one or becomes a stop codon which functions as a sign of translation-termination. The former case is called a *silent site mutation (non-synonymous mutation)* and the latter one is a *replacement site mutation (synonymous mutation)*. The majority of mutations at the first and second positions of a codon are replacement mutations while mutations happened at the third codon position usually result in silent mutations. For example, UUU is a codon for encoding amino acid phenylalanine. If the first U is replaced by C, then CUU encodes leucine but UUC still codes for the same amino acid as UUU does.

Mutations will give rise to alternate forms of genes, called *alleles* (Raineri (2001)). The wild type allele refers to the nucleotide sequence of a gene which is characteristic of most individuals of a species. It is also called the normal, standard, or reference allele and usually designated by capital letter. The mutant type is represented by lower case letter. While silent mutations have no impact on the fitness of neither an individual nor its descendants due to the fact that they will not alter the structure of the polypeptide chain, the severity of the effect of replacement mutations may vary case by case. Most of the replacement mutations lead to critical changes in protein structure and functionality

or even completely inactivate protein producing process and thus are detrimental to the host, but a significant part shows mild effect and can be beneficial. The effect of mutations at a gene is quantified by the *fitness coefficient* of the gene, ω, defined by the expected relative number of living descendants of that gene in next generation (Ewens (2003)). In a simplified situation where a gene is expressed by a single allele, we use A for the wild type type allele and a the mutant. The fitness coefficient of the gene is define as $\omega(A) = 1$ and $\omega(a) = 1 + \sigma$. In population genetics, for a haploid population of size N the parameter σ can be scaled in terms of the effective population size N such that $\sigma \sim \frac{\gamma}{N}$ for large N and the γ is called the *selection coefficient* of the gene. Apparently, a positive γ indicates an advantageous mutation and a negative γ means that the mutation is deleterious. Neutral mutations are usually assumed to have zero γ values.

The field of population genetics is concerned with identifying and quantifying the role of genetic forces such as mutation, selective effects and demographic forces in an evolution history of certain genes in particular populations (Lewontin (1974)). Consider a single DNA site in one species. If there is only one type of nucleotide at that site across all individuals in a population, we say that the site is *fixed* or *monomorphic* at that nucleotide. In contrast, the site is *polymorphic* if there exists more that one type of nucleotide among different individuals. For a random sample come from two species which are close relatives, such as humans and chimpanzees or the *Drosophila* species *melanogaster* and *simulans*, if a DNA site is fixed at one nucleotide within one species but the homologous site is fixed at a different nucleotide within the other species, then we say that the site is a *fixed difference*. While if the site is polymorphic within at least one of the two species, it is a *polymorphism*.

With data collected from only one species, the corresponding statistical analysis is usually not powerful enough to detect the selective effect unless the polymorphism is substantial (Sawyer and Hartl (1992)). Instead, since almost all species in existence have arisen in evolution from ancestral species and somehow related, approaches comparing site polymorphism and differences within and between two closely related daughter species can be useful. We consider nucleotide sequences from a certain genetic locus and a random sample of n_1 individuals that is chosen from one species and another sample of n_2 individuals chosen from the other species. We assume that the two species have diverged from their most recent common ancestor at a time t in the past. The information gained can be seen as a $n_1 + n_2$ by l matrix of elements consisting of the four nucleotides A, T, C, G if the sequences are aligned by corresponding DNA sites with the length of each DNA sequence being l. We shall see no more than two letters in any column since there are at most two types of nucleotides in the two populations under the assumption that mutations are rare enough so they will occur at most once at a site. McDonald and Kreitman (1991) proposed a statistical test of neutrality utilizing the following two way contingency table in which the cell counts are summarized from the matrix described above,

	D	P
S	F_s	V_s
R	F_r	V_r

Table 1.1. DPRS Table

Here F_s and F_r represent the total numbers of silent and replacement fixed differences

whilst V_s and V_r denotes the numbers of silent and replacement polymorphisms, respectively. This table is called McDonald-Kreitman table or DPRS table, where D is abbreviation for fixed difference, P is for polymorphism, R is for replacement and S is for silent.

One drawback of the McDonald-Kreitman table is its inability to distinguish legacy polymorphisms from new polymorphisms (Amei and Sawyer (2010)). The legacy polymorphism refers to a polymorphic site at which the mutation occured before the daughter species diverged from their common ancestor while the new polymorphism is a polymorphic site at which the mutation occured in one of the two daughter populations after the divergence of the two species. Recall that we have assumed that mutations are sufficiently rare so that once an alternation occurs we shall never see another at the same site. Thus the feature of new polymorphism implies that the corresponding sites can only be polymorphic in at most one sample of the two populations. However, legacy polymorphism may result in sites that are simultaneously polymorphic in both population samples. Furthermore, due to the low starting population frequency of a new mutation ($\frac{1}{N} \approx 0$ for large N, where N is the effective population size), most mutant nucleotides at new polymorphic sites can not survive through reproduction under random mating assumption, especially for the first several generations after the mutation. In other words, for recently diverged two species, we have a relatively higher chance to see sites that are polymorphic in both samples that are descendants of legacy polymorphic sites.

The facts described above are used to extend the McDonald-Kreitman table so that the accuracy of estimating population parameters can be improved. The table shown below is such an extension, designated by DOHRS table, where W_s, W_r are the numbers

	D	O	H
S	F_s	O_s	B_s
R	F_r	O_r	B_r

Table 1.2. DOHRS Table

of sites that are polymorphic in only one sample and B_s, B_r are the numbers of sites that are polymorphic in both samples. This classification naturally implies $V_s = O_s + B_s, V_r = O_r + B_r$. To be clear and concise, we name a site as a *F-site* if it is a fixed difference, a *O-site* if it is polymorphic in only one sample and a *B-site* if it is polymorphic in both samples.

Under the assumption of neutral selective effect ($\gamma = 0$) on silent mutations, the null hypothesis of the McDonald and Kreitman test (MK test) is that the ratio of replacement to silent polymorphisms within species should be equal to the ratio of replacement to silent fixed differences between species, i.e. $\frac{V_r}{V_s} = \frac{K_r}{K_s}$. The test was validated using DNA sequences data of the *alcohol dehydrogenase gene (Adh)* locus in three species of the *Drosophila melanogaster* species subgroup and showed an excess number of replacement fixed differences which provided an evidence of positive selection. However the MK test does not take into account the information of allele frequency spectrum of polymorphisms and hence lacks the power in detecting the strength and direction of selection (Sawyer et al. (2007),Williamson et al. (2005)).

Sawyer and Hartl (1992) proposed a Poisson Random Field (PRF) model and showed that under certain biological assumptions the distributions of those counts in Table 1.1 can be modeled as Poisson random variables with means depending on the popu-

lation level site frequencies of mutant nucleotide. They showed that the means of those Poisson counts are functions of genetic parameters such as speciation time, mutation rate and selection coefficient and with proper sampling formulas, one will be able to make statistical inference about selection and divergence of two closely related biological species (See also Hartl et al. (1994),Sawyer (1994),Akashi (1999),Bustamante et al. (2001), Amei and Sawyer (2012)). Contribution to multiple loci analysis had been made by Bustamante et al. (2002). They designed a hierarchical Baysian model based on the PRF theory and implemented it using Markov chain Monte Carlo (MCMC) simulation to estimate various genetic parameters of interest.

Although the original PRF model had proposed a powerful and attractive approach to analyzing and interpreting DNA site polymorphism within and between species, there still exist potential improvements of the model by relaxing or even removing some artificial biological assumptions. One common criticism on the model is that alleles at different loci are assumed at *linkage equilibrium* or free (in high level) of recombination, which translates into the independence of nucleotide sites. Also for mathematical simplicity, the distribution of site frequency spectrum is taken to be stationary after the divergence of two species and thus the model is called a *time-independent (time-equilibrium, time-homogeneous)* PRF model. Besides, the selective effects of replacement mutations within any particular gene is assumed to be fixed and only varies from one gene to another. In other words, γ remains a constant for each gene and this is called *fixed effect* assumption. The original formulation also stipulates that another important factor impacting on changes in mutant site frequency, i.e. the effective population size, stays as a constant for both ancestral and daughter species involved in the analysis.

Many authors have put effort on refining the basic PRF model with more general biological settings. For instance, the likelihood ratio test (Hartl et al. (1994)) derived from the PRF model lacks robustness against departures from the linkage equilibrium assumption (Bustamante et al. (2001)). Zhu and Bustamante (2005) then proposed a composite likelihood method to correct the bias of estimates using simulations with a specified recombination rate. For relaxing the fixed effect assumption, Sawyer et al. (2003) proposed a random effect model in which the selective intensity of arising mutations within a given genetic locus was assumed to have Gaussian distribution with mean (but not variance) changing between genes. Additional inference on a dataset of 91 genes in two *Drosophila* species, *D.melanogaster* and *D.simulans*, was presented in Sawyer et al. (2007) using this random effect model. Later, simulation studies suggested that the divergence time tended to be overestimated with time-equilibrium models (Abel (2009)). Amei and Sawyer (2010) then removed the assumption and built a time-dependent PRF model by explicitly accounting for the time since the divergence of the two species into the model by using diffusion approximation to discrete time discrete state Markov chains. A strong precision of their model was verified on various simulated datasets and the application of the model to the 91 *Drosophila* genes data yielded a more accurate estimate of divergence time compared to the results of time-independent models (Amei and Sawyer (2012)). However, the restrictive fixed effect setting was still preserved in Amei and Sawyer's time-dependent model. More recently, Zhou (2013) elaborated a sophisticated time-dependent random effect framework simultaneously considering within-locus randomness of selective effect and mutation-selection-drift disequilibrium after divergence.

However all models mentioned above ignored the change in effective population size

over evolution process which can be a confounding factor in inferring natural selection on DNA polymorphism patterns. Williamson et al. (2005) presented a population size change model to simultaneously make statistical inference of selection and demographic factor. Their model assumes that a species experienced an abrupt change from the ancestral population size to current size at some moment during the evolution. A maximum likelihood approach incorporating the ratio of the two population sizes was applied to a large dataset of 301 sequenced human genes collected from 90 individuals and the results revealed strong proof of population expansion under common negative selection on replacement mutations. Boyko et al. (2008) extended Williamson's approach to infer the distribution of fitness effects given a non-stationary demographic history. However, their studies are based on site frequency spectrum data from single population and simulation results have shown that frequency spectrum polymorphism data may generate strongly biased estimates of selection parameters even for minor deviation from the model assumption of genic selection (Williamson et al. (2004)). Gutenkunst et al. (2009) proposed a diffusion-based method to compare different demographic models based on the joint distribution of allele frequencies in multiple populations, but their model did not target at inferring other genetic parameters such as mutation rate and selection coefficient.

In this dissertation, we explore a modified time-dependent Poisson Random Filed model to depict the DNA site polymorphisms within and between two closely associated species while accounting for the differences in their population sizes compared to that of the ancestral species. We begin with a comprehensive set up of the time-dependent PRF theory in Chapter 2. A non-constant population size time-inhomogeneous PRF model and sampling formulas for multi-loci data are then presented in Chapter 3. A hierarchical

Baysian implementation of the model is introduced in the first part of Chapter 4. Numerical evaluation of diffusion equations involved in our model is extremely computational demanding and hence forces us to develop an efficient program. The related technique details of linking R code with `C++` as well as using a parallel scheme constitute the second part of Chapter 4. Results of simulation studies and application to the 91 *Drosophila* genes data mentioned above are discussed in Chapter 5 together with a second thought on our setting given the discovered dependence between biased estimates of divergence times and population size ratios.

CHAPTER 2

POISSON RANDOM FIELD MODEL FOR SELECTION

2.1 Poisson Random Field

Methods of using stochastic processes to model the change of gene frequencies over time have been developed for decades. A random field is a generalized stochastic process, usually taking values in a Euclidean space and defined over a parameter space of dimensionality at least one (Adler and Taylor (2007)). We start with the definition of *Poisson Random Field (PRF)* as follows.

Definition 2.1. A random measure $(X, \mathcal{F}, \boldsymbol{N})$ on a measurable space $(X, \mathcal{F})$ with *mean measure* $(X, \mathcal{F}, \mu)$ is a *Poisson Random Field* if

$$E\left(\exp\left(\int_X f(x)\boldsymbol{N}(dx)\right)\right) = \exp\left(\int_X \left(e^{f(x)} - 1\right)\mu(dx)\right) \tag{2.1}$$

for all bounded $\mathcal{F}$-measurable functions $f(x)$ on X with $\int_X |f(x)|\mu(dx) < \infty$, where for $\forall\ A \subseteq X$ $\boldsymbol{N}(A)$ are random variables such that $(X, \mathcal{F}, \boldsymbol{N})$ is a measure with probability one.

Now consider a special case of (2.1) on the set $X = \{0, 1, \ldots, n\}$. We use a vector $\boldsymbol{N} = (N_0, N_1, \ldots, N_n)$ of length $n+1$ to define a random measure $\boldsymbol{N}(A) = \sum_{i\in A} N_i$ for $A \subseteq X$, where $N_i's$ are independent Poisson random variables with means $E(N_i) = u_i$. Next let μ be a measure such that $\mu(A) = \sum_{i\in A} u_i$, thus $E[\boldsymbol{N}(A)] = \sum_{i\in A} u_i = \mu(A)$. It

can be shown that

$$E\left(\exp\left(\sum_{i=0}^{n} c_i N_i\right)\right) = \exp\left(\sum_{i=0}^{n} u_i \left(e^{c_i} - 1\right)\right) \tag{2.2}$$

for any numbers c_i, $0 \leq i \leq n$. On the other side, if (2.2) is true for any set of c_i's, given $\boldsymbol{N} = (N_0, N_1, \ldots, N_n)$ is an arbitrary set of $n+1$ random variables, then the N_i's are independent Poisson random variables with means $E(N_i) = u_i$. In the following part of this section, all results involving the PRF refer to the case defined above unless explicitly specified.

Poisson random variables are frequently used in counting processes. Thus $\boldsymbol{N} = (N_0, N_1, \ldots, N_n)$ can be treated as a vector of counts, in which N_i represents the number of objects at state i. Suppose that at a specific moment, all objects jump to some states in a finite state space Y and transitions are mutually independent. For any object initially at state i, it jumps to state $j, j \in Y$ with probability p_{ij} and $\sum_{j \in Y} p_{ij} = 1$. Denote Q_j as the count of objects at state $j \in Y$ after jumping. Then it has been shown that (Amei and Sawyer (2010))

Lemma 2.1. $\mathbf{Q} = \{Q_j : j \in Y\}$ *is a set of independent Poisson random variables with means* $E(Q_j) = \sum_{i \in X} u_i p_{ij}$.

If we simply let $Y = X = \{0, 1, \ldots, N\}$ and assume that transitions occur at discrete time, then those objects actually move in a manner of independent time-homogenous Markov chains according to the transition probabilities p_{ij}. The way of modelling is inspired by the structure of chromosomes, which can be seen as strings of letters composed

of nucleotides. For a haploid population of size N, suppose we have an alignment of DNA sequences in which the rows are sequences for each individual and the columns are corresponding sites of nucleotides. Objects mentioned above are in fact sites on chromosomes. For a particular site, the state $i \in \{0, 1, \ldots, N\}$ refers to the number of mutant nucleotides in the population, or equivalently, $\frac{i}{N} \in S_N = \{0, \frac{1}{N}, \ldots, \frac{N-1}{N}, 1\}$ refers to the mutant frequency. The transitions described in Lemma 2.1 can be compared to reproduction of individuals, but the source of change of gene frequencies also includes mutation. So we need to extend the model further to cover possible arrivals of new mutants.

Consider a PRF purely formed by immigrants, i.e., there is no pre-existing objects at any state. Define R_l as the number of new immigrants joining the system at time step $l \geq 1$, which are assumed to be identical Poisson random variables with mean μ. Due to the fact that mutations are so rare at site level that chance of multiple mutations at the same site can be ignored, it will be reasonable to assign 1 as the initial state for all immigrants. We assume that all objects in the system will move one step forward independently according to p_{ij}. Let $N_{i,k}$ be the number of objects at state i at step k, which is counted right after the arrival of new immigrants but before any further transition occurs. Using Lemma 2.1, it can be shown that (Amei and Sawyer (2010))

Lemma 2.2. *At any step $k \geq 1$, $\{N_{i,k}\}$ is a set of independent Poisson random variables with means $E(N_{i,k}) = \mu \sum_{l=0}^{k} p_{1i}^{(k-l)}, \qquad 0 \leq i \leq n,$*

where $\mathbf{P}^{(k-l)} = [p_{ij}^{(k-l)}]_{(N+1)\times(N+1)}$ is the $(k-l)$th power of the transition matrix $\mathbf{P} = [p_{ij}]_{(N+1)\times(N+1)}$, in particular, $\mathbf{P}^{(0)} = \mathbf{I}$. So $p_{1i}^{(k-l)}$ gives the transition probability that an

object with starting state 1 arrives state i using $k-l$ steps.

2.2 Diffusion Approximation and Population results

We will apply the Moran's second model to derive the transition probabilities p_{ij} for mutant nucleotide frequency (Moran (1959)). It involves generation overlap in the sense that an individual who died at a specific moment will be replaced by a randomly chosen member in the population before the death (can be the same one). Hence the population remains fixed at the haploid size of N and one generation is equivalent to N discrete time steps. Let X_k be the number of mutant individuals at kth moment, since exactly one individual will be replaced at each time step, transition of the states will only occur in the following three ways: $j \to j-1, j \to j+1, j \to j$. Without selection, transition probabilities are trivial:

$$\begin{aligned} p_{j,j+1} = p_{j,j-1} &= \frac{N-j}{N} \cdot \frac{j}{N}, \\ p_{j,j} &= 1 - p_{j,j+1} - p_{j,j-1}, \qquad 1 \le j \le N-1, \\ p_{0,0} &= p_{N,N} = 1. \end{aligned}$$

Given the fitness coefficient $\omega = 1+\sigma$, then for $1 \le j \le N-1$,

$$\begin{aligned} p_{j,j+1} &= \frac{N-j}{N} \frac{(1+\sigma)j}{(1+\sigma)j + (N-j)}, \\ p_{j,j-1} &= \frac{j}{N} \frac{N-j}{(1+\sigma)j + (N-j)}, \\ p_{j,j} &= 1 - p_{j,j+1} - p_{j,j-1}. \end{aligned} \tag{2.3}$$

Notice that 0 and 1 are traps or absorbing states since $p_{0,0} = p_{N,N} = 1$, which represent the loss of the mutant allele or its fixation at a site.

Now consider a model including both the mutations happened before the divergence of the two species and after the divergence. Assume that there are R_0 population polymorphic sites at time step 0 and R_l sites with new mutation appeared at time step $l \geq 1$. Under no repeated mutations assumption, sites represented by R_0 and R_k are distinct from each other and can be modeled as Poisson random variables with mean $E(R_0) = \mu_0, E(R_k) = \mu$. As mentioned before, the frequencies of mutant nucleotide at those sites will follow independent Markov chains with transition probabilities given in (2.3) if we assume that nucleotide sites evolve independently which is equivalent to assume that sites are at linkage equilibrium. Processes recording the related frequencies are denoted by $\{X_{a,k}\}(1 \leq a \leq R_0, k \geq 1)$ for initial polymorphic sites and $\{Y_{b,l,k}\}(1 \leq b \leq R_l, 1 \leq l \leq k, k \geq 1)$ for sites with new mutation at time step l. Notice that the sample spaces of the two Markov chains $X_{a,k}$ and $Y_{b,l,k}$ are $S_N = \{0, \frac{1}{N}, \ldots, \frac{N-1}{N}, 1\}$ and $Y_{b,l,l} = \frac{1}{N}$ for any $l \geq 1$.

The above described Markov chains are based on discrete time and discrete state space. Whereas, in real world case, the population size of existing species and their divergence time from ancestral species are large enough to be considered infinite. Using diffusion approximation to discrete time discrete state Markov chains, it has been shown that the processes $X_{a,k}$ and $Y_{b,l,k}$ converge in distribution to a diffusion process X_t on the open interval $(0, 1)$ with 0 and 1 being absorbing states and the diffusion time t is scaled in units of N generations ($\frac{k}{N^2} \to t$, since 1 generation equals N discrete Moran time steps; see Amei and Sawyer (2010)).

Specifically, given the following parameters scaled by the population size N,

$$N\mu \to \theta,\ N\sigma \to \gamma,\ \frac{k}{N^2} \to t, \quad \text{when } N \to \infty,$$

they presented the distribution of mutant site polymorphisms and expected sites of fixation of mutants as follows.

Theorem 2.1. *For any function $f(x)$ continuous on (0,1) with $f(0) = f(1) = 0$,*

$$\begin{aligned}
&\lim_{N\to\infty} E\left(\sum_{j=1}^{N-1} f(\frac{j}{N})N_{j,k}\right) \\
&\quad = \lim_{N\to\infty}\left(\sum_{a=1}^{R_0} f(X_{a,k}) + \sum_{l=1}^{k}\sum_{b=1}^{R_l} f(Y_{b,l,k})\right) \\
&\quad = \int_0^1\int_0^1 p(t,x,y)f(y)dm(y)d\delta(x) \\
&\quad +\theta\int_0^1 \frac{s(1)-s(x)}{s(1)-s(0)}(f(x) - \int_0^1 p(t,x,y)f(y)dm(y))dm(x),
\end{aligned} \tag{2.4}$$

where

$$s(x) = \frac{1-e^{-\gamma x}}{\gamma}\ ;\ dm(x) = \frac{e^{\gamma x}}{x(1-x)}dx$$

are referred to as the scale function and speed measure of the corresponding diffusion process (Karlin and Taylor (1981)). The $p(t,x,y)$ in 2.4 is a transition density function which is symmetric in the sense that $p(t,x,y) = p(t,y,x)$ and $d\delta(x)$ is the limiting distribution of site frequencies at diffusion time 0, which we assume to be a Borel measure defined on (0,1) (Amei and Sawyer (2010)).

Theorem 2.2. *Given the assumptions in Theorem 2.1, the limiting expected number of sites that have been saturated by the mutant nucleotide on diffusion time $(0,t]$ is*

$$\begin{aligned}\lim_{N\to\infty} E(N_{N,k}) &= \lim_{N\to\infty} E\left(\sum_{i=1}^{N-1} u_i p_{iN}^{(k)} + \mu \sum_{l=0}^{k-1} p_{1N}^{(l)}\right) \\ &= \int_0^1 P_x(T_1 \le t) d\delta(x) + \frac{\theta}{s(1)} \int_0^t \widetilde{P}_0(T_1 \le u) du \\ &= \frac{1}{s(1)} \left(\int_0^1 s(x) d\delta(x) - \int_0^1 \int_0^1 p(t,x,y) s(y) dm(y) d\delta(x) \right. \\ &\quad \left. + \theta t - \theta \int_0^1 \int_0^1 q(u,0+,y) s(y)^2 dm(y) du \right)\end{aligned} \tag{2.5}$$

where $T_y = \inf\{t \ge 0 : X_t = y\}$ *represents the hitting time of mutant frequency* y, $\widetilde{P}_0(T_1 \le u) = P_0(T_1 \le u | T_1 < T_0)$ *corresponds to the dual process of* X_t, $\widetilde{X}_t = X_t | T_1 < T_0$, *with* $q(t,x,y) = \tilde{p}(t,x,y) = \dfrac{p(t,x,y)}{s(x)s(y)}$ *as its transition density*[*].

Theorem 2.1 gives the weighted expected number of sites that are population polymorphic at diffusion time t, from which we can derive the limiting density of the PRF as

$$\begin{aligned}\lambda(y|\theta,\gamma,t) = \Bigg[&\int_0^1 p(t,x,y) d\delta(x) + \theta \frac{s(1)-s(y)}{s(1)-s(0)} \\ &- \theta \int_0^1 \frac{s(1)-s(x)}{s(1)-s(0)} p(t,x,y) dm(x) \Bigg] dm(y)\end{aligned} \tag{2.6}$$

In other words, one can get the expected number of sites with mutant frequency $y \in (y_1, y_2) \subset (0,1)$ by integrating (2.6) over the range (y_1, y_2).

Since it is impractical to draw information from all individuals in the target species,

[*] $q(u,0+,y) = \lim_{x\to 0+} q(u,x,y)$

people use the above population level results to derive sampling formulas to calculate means of the Poisson random variables $F_s, O_s, B_s, F_r, O_r, B_r$ (Amei and Sawyer (2010)). Then the parameters such as the speciation time t, selection coefficient γ and mutation rate θ can be analyzed through likelihood functions involving the observed counts.

In the above time-dependent PRF model, one of the key assumptions is the constant population size N across the ancestral and two daughter species, which suggests the divergence times t, the mutation rates θ and the selection coefficients γ are the same for daughter species. In next chapter, we will loose this biologically unrealistic assumption by proposing a non-constant population size model and also present corresponding sampling formulas.

CHAPTER 3

A PRF MODEL WITH NON-CONSTANT POPULATION SIZE

3.1 Construction of The Model

We denote the effective population size of the ancestor species as N_a and those of the two daughter species as N_1 and N_2, respectively. Suppose that the population sizes of the two daughter species are proportional to that of the ancestral species, that is

$$\nu_1 = \frac{N_1}{N_a}, \qquad \nu_2 = \frac{N_2}{N_a}, \tag{3.1}$$

We also assume that the selection coefficient and the mutation rate at a legacy site satisfies $N_a\sigma \sim \gamma$ and $N_a\mu \sim \theta$ for large N_a. Then the mutation rates and the selection coefficients at a single genetic locus in the two daughter species can be derived as

$$N_1\mu = \nu_1 N_a\mu \sim \nu_1\theta, \quad N_2\mu = \nu_2 N_a\mu \sim \nu_2\theta,$$

and

$$N_1\sigma = \nu_1 N_a\sigma \sim \nu_1\gamma, \quad N_1\sigma = \nu_2 N_a\sigma \sim \nu_2\gamma, \quad \text{for large } N_a$$

In the above expressions, μ is the aggregated nucleotide mutation rate per generation and σ is the same selection coefficient per generation. The γ and θ are then called the scaled selection coefficient and the scaled mutation rate.

Under the assumption of independent Moran time steps* k_1 and k_2, the diffusion

*Discussion of this assumption will be presented in Chapter 5.

times t_1 and t_2 satisfy

$$\frac{k_1}{N_1^2} \to t_1, \qquad \frac{k_2}{N_2^2} \to t_2$$

as $N_1, N_2 \to \infty$ respectively. Let $\boldsymbol{\alpha} = \boldsymbol{\alpha_s} = (t_1, t_2, \theta_s, \nu_1, \nu_2)$ be a set of parameters corresponding to a silent site mutation and $\boldsymbol{\alpha} = \boldsymbol{\alpha_r} = (t_1, t_2, \gamma, \theta_r, \nu_1, \nu_2)$ be a set of parameters related to a replacement site mutation. Given a DNA alignment of n_1 sequenced genes from a single genetic locus in one species and n_2 sequences from the orthologous gene in a related species, the FOB counts in a single DOHRS table remain Poisson distributed with means calculated by formulas depending on $\boldsymbol{\alpha_s}$ and $\boldsymbol{\alpha_r}$ (Amei and Sawyer (2010)). Each site counted in the sample is either a legacy polymorphic site or a new polymorphic site and hence we need to measure the contribution made by the two types of polymorphic sites individually. Meanwhile, a site that is monomorphic in a sample can be due to a fixation at this site in the population or just due to random draws. We will take this fact into account when we develop corresponding sampling formulas.

3.2 Contribution of Legacy Polymorphic Sites

By Theorem 2.1 and the first term in 2.6, the distribution of polymorphic site frequencies in daughter species i $(i = 1, 2)$ at diffusion time t_i that are derived from legacy polymorphic sites is a PRF with mean density

$$\lambda_{LP}(y|\boldsymbol{\alpha}) = \int_0^1 p(t_i, x, y) d\delta(x),$$

Using Theorem 2.2, the expected number of sites that have become fixed at mutant nucleotides and are descendants of legacy polymorphic sites is Poisson with mean

$$H_{LP}(\boldsymbol{\alpha}) = \int_0^1 P_x(T_{1,i} \leq t_i) d\delta(x),$$

where $T_{y,i}$ is the hitting time of mutant frequency y for species i. Notice that the equilibrium distribution $d\delta(x) = \theta \frac{s(1)-s(x)}{s(1)-s(0)} dm(x)$ appeared in $\lambda_{LP}(y|\boldsymbol{\alpha})$ and $H_{LP}(\boldsymbol{\alpha})$ is free of the genetic parameters corresponding to the daughter species since it is the mean density of mutant frequencies in the ancestral species.

Suppose that a legacy polymorphic site in species i has not become fixed at either of the two alleles at t_i. In other words, the population mutant frequency at the site is $y \in (0, 1)$. For a random sample of size n_i, the probability that the site is monomorphic at the wild type nucleotide is $(1-y_i)^{n_i}$, the probability that it is monomorphic at the mutant nucleotide is y^{n_i} and the probability that the site contains both types of nucleotides is $1 - y^{n_i} - (1 - y_i)^{n_i}$.

For an arbitrary legacy polymorphic site with starting mutant frequency of $x \in (0, 1)$, we use $g_1(n_i, x)$ to denote the probability that only the non-mutant nucleotide shows up in the sample, $g_2(n_i, x)$ to denote the probability that the site is polymorphic in the sample and $g_3(n_i, x)$ the probability that only the mutant nucleotide is present in the

sample. Then these probabilities are given specifically as follows

$$
\begin{aligned}
g_1(x,n_i) &= P_x(T_{0,i} < t_i) + \int_0^1 p(t_i,x,y)(1-y)^{n_i} dm_i(y) \\
&= 1 - \frac{s_i(x)}{s_i(1)} - \int_0^1 p(t_i,x,y)\left[1-(1-y)^{n_i} - \frac{s_i(y)}{s_i(1)}\right] dm_i(y) \\
g_2(x,n_i) &= \int_0^1 p(t_i,x,y)[1-y^{n_i}-(1-y)^{n_i}] dm_i(y) \\
g_3(x,n_i) &= P_x(T_{1,i} < t_i) + \int_0^1 p(t_i,x,y) y^{n_i} dm_i(y) \\
&= \frac{s_i(x)}{s_i(1)} + \int_0^1 p(t_i,x,y)\left[y^{n_i} - \frac{s_i(y)}{s_i(1)}\right] dm_i(y)
\end{aligned}
\tag{3.2}
$$

Recall that for a single nucleotide site in a population of size N, the change of mutant frequency is modeled by the discrete time and discrete state space Markov chains $X_{a,k}$ and $Y_{b,l,k}$ introduced in Chapter 2. For daughter species i, the state space of those Markov chains depends on the effective population size N_i and becomes $S_{N_i} = \{0, \frac{1}{N_i}, \ldots, \frac{N_i-1}{N_i}, 1\}$ instead of S_N under constant population size assumption. With the modified genetic parameters, chains modelling daughter species i converges weakly to the diffusion process X_{t_i} on state space (0,1) with speed measure $dm_i(y) = \dfrac{e^{\nu_i\gamma y}}{y(1-y)}dy$ and scale function $s_i(y) = \dfrac{1-e^{-\nu_i\gamma y}}{\nu_i\gamma}$. Hence population size ratios ν_1 and ν_2 are implicitly implemented into our model through $dm_i(y)$ and $s_i(y), i = 1, 2$.

Given a joint DNA sequence alignment sample of size $n_1 + n_2$ from two closely related species as described previously, we denote $P_1(x)$ as the probability that a legacy polymorphic site becomes a F-site in the sample, $P_2(x)$ as the probability that it becomes a

O-site in the sample and $P_3(x)$ the probability that it becomes a B-site. Then

$$\begin{aligned} P_1(x) &= g_1(x,n_1)g_3(x,n_2) + g_1(x,n_2)g_3(x,n_1) \\ P_2(x) &= g_2(x,n_1)\left[g_1(x,n_2)+g_3(x,n_2)\right] + g_2(x,n_2)\left[g_1(x,n_1)+g_3(x,n_1)\right] \\ &= g_2(x,n_1) + g_2(x,n_2) - 2g_2(x,n_1)g_2(x,n_2) \\ P_3(x) &= g_2(x,n_1)g_2(x,n_2) \end{aligned} \tag{3.3}$$

Let L_1, L_2, L_3 be the number of $F-, O-, B-$ sites in the sample which are derived from legacy polymorphic sites, respectively. Then $L_j (j = 1, 2, 3)$ are Poisson random variables and integrating $P_j(x)$ over the equilibrium density $d\delta(x)$ yields the means of the $L_j (j = 1, 2, 3)$. We denote these means using $\Psi_{LP,j}(\boldsymbol{\alpha}, n_1, n_2)$, then

$$\begin{aligned} E(L_j) =& \Psi_{LP,j}(\boldsymbol{\alpha}, n_1, n_2) \\ &= \int_0^1 P_j(x) d\delta(x) \\ &= \theta \int_0^1 P_j(x) \frac{s(1)-s(x)}{s(1)-s(0)} dm(x), \quad j = 1, 2, 3. \end{aligned} \tag{3.4}$$

3.3 Contribution of New Polymorphic Sites

For daughter species $i, i = 1, 2$, we can measure the proportion of new polymorphic sites that are polymorphic at time t_i and obtain the mean density of the corresponding Poisson Random Field by the remaining terms of (2.6) except the first one, that is

$$\begin{aligned} \lambda_{NP}(y|\boldsymbol{\alpha}) &= \nu_i\theta \frac{s_i(1)-s_i(y)}{s_i(1)-s_i(0)} - \nu_i\theta \int_0^1 \frac{s_i(1)-s_i(x)}{s_i(1)-s_i(0)} p(t_i,x,y) dm_i(x) \\ &= \theta \cdot \frac{\nu_i}{s_i(1)} \left[s_i(1) - s_i(y) - \int_0^1 (s_i(1)-s_i(x)) p(t_i,x,y) dm_i(x) \right] \end{aligned} \tag{3.5}$$

Since we have assumed no repeated mutations will occur at the same site, a new polymorphic site will be classified as a F-site in the $n_1 + n_2$ joint sample if either the site has become fixed at the mutant nucleotide in species i during $(0, t_i]$ or by chance the sample of species i only has the mutant nucleotide (i.e., n_i draws) although the site has not become fixed in the population of species i. The expected number of mutant fixations in species i at time t_i is a Poisson random variable. Its mean can be obtained by slightly modifying the second part of Theorem 2.2, i.e.,

$$\begin{aligned} &\frac{\nu_i \theta}{s_i(1)} \int_0^1 \tilde{P}(T_{1,i} < u) du \\ &= \theta \cdot \frac{\nu_i}{s_i(1)} \left[t_i - \int_0^{t_i} q_i(u, 0+, y) s_i^2(y) dm_i(y) du \right], \end{aligned} \tag{3.6}$$

where $q_i(u, 0+, y) = \lim_{x \to 0+} \dfrac{p(t_i, x, y)}{s_i(x) s_i(y)}$.

Let $\Psi_{NP,1}(\boldsymbol{\alpha}, n_i)$ be the expected number of sites that are monomorphic in the sample of species i,

$$\Psi_{NP,1}(\boldsymbol{\alpha}, n_i) = \frac{\nu_i \theta}{s_i(1)} \int_0^1 \tilde{P}(T_{1,i} < u) du + \int_0^1 \lambda_{NP}(y|\boldsymbol{\alpha}) y^{n_i} dm_i(y) \tag{3.7}$$

Similarly, the number of polymorphic sites in the sample of species i at time t_i due to new polymorphism is Poisson distributed with the following mean, which we label as $\Psi_{NP,2}(\boldsymbol{\alpha}, n_i)$,

$$\Psi_{NP,2}(\boldsymbol{\alpha}, n_i) = \int_0^1 \lambda_{NP}(y|\boldsymbol{\alpha})(1 - y^{n_i} - (1-y)^{n_i}) dm_i(y) \tag{3.8}$$

These sites make up the new polymorphism contribution to O-sites. And as mentioned earlier, any B-site can not be an offspring of a new polymorphism.

3.4 Sampling Formulas for The FOB Counts

Combining the contributions from the legacy and the new polymorphic sites together gives the following theorem of the sampling formulas.

Theorem 3.1. *For two daughter species with the effective population sizes N_1 and N_2 scaled as in (3.1), the sample counts $F_s, O_s, B_s, F_r, O_r, B_r$ in Table 1.2 are independently distributed Poisson random variables with means*

$$\begin{aligned}
E(F_s) &= \Psi_{LP,1}(\boldsymbol{\alpha_s}, n_1, n_2) + \Psi_{NP,1}(\boldsymbol{\alpha_s}, n_1) + \Psi_{NP,1}(\boldsymbol{\alpha_s}, n_2) \\
E(O_s) &= \Psi_{LP,2}(\boldsymbol{\alpha_s}, n_1, n_2) + \Psi_{NP,2}(\boldsymbol{\alpha_s}, n_1) + \Psi_{NP,2}(\boldsymbol{\alpha_s}, n_2) \\
E(B_s) &= \Psi_{LP,3}(\boldsymbol{\alpha_s}, n_1, n_2) \\
E(F_r) &= \Psi_{LP,1}(\boldsymbol{\alpha_r}, n_1, n_2) + \Psi_{NP,1}(\boldsymbol{\alpha_r}, n_1) + \Psi_{NP,1}(\boldsymbol{\alpha_r}, n_2) \\
E(O_r) &= \Psi_{LP,2}(\boldsymbol{\alpha_r}, n_1, n_2) + \Psi_{NP,2}(\boldsymbol{\alpha_r}, n_1) + \Psi_{NP,2}(\boldsymbol{\alpha_r}, n_2) \\
E(B_r) &= \Psi_{LP,3}(\boldsymbol{\alpha_r}, n_1, n_2)
\end{aligned} \tag{3.9}$$

Notice that all components of $\Psi_{LP,j}(j = 1, 2, 3)$ and $\Psi_{NP,j}(j = 1, 2)$ contain the scaled mutation rate θ_s for a silent site and θ_r for a replacement site, we rewrite the means in the following format uniformly for a synonymous or non-synonymous sites as follows

$$\begin{aligned}
E(F) &= \theta\Lambda_1(t_1, t_2, \nu_1, \nu_2, \gamma, n_1, n_2) \\
E(O) &= \theta\Lambda_2(t_1, t_2, \nu_1, \nu_2, \gamma, n_1, n_2) \\
E(B) &= \theta\Lambda_3(t_1, t_2, \nu_1, \nu_2, \gamma, n_1, n_2),
\end{aligned} \tag{3.10}$$

in which

$$
\begin{aligned}
\Lambda_1 &= \frac{1}{s(1)} \int_0^1 \left[g_1(x,n1) g_3(x,n2) + g_1(x,n2) g_3(x,n1) \right] (s(1) - s(x))\, dm(x) \\
&\quad + \sum_{i=1,2} \frac{\nu_i}{s_i(1)} \left[t_i - \int_0^{t_i} \int_0^1 q_i(u, 0+, y) s_i^2(y) dm_i(y) du \right] \\
&\quad + \sum_{i=1,2} \frac{\nu_i}{s_i(1)} \int_0^1 \left[s_i(1) - s_i(y) - \int_0^1 (s_i(1) - s_i(x)) p(t_i, x, y) dm_i(x) \right] y^{n_i} dm_i(y) \\
\Lambda_2 &= \frac{1}{s(1)} \int_0^1 \left[g_2(x,n_1) + g_2(x,n_2) - 2 g_2(x,n_1) g_2(x,n_2) \right] (s(1) - s(x))\, dm(x) \\
&\quad + \sum_{i=1,2} \frac{\nu_i}{s_i(1)} \left[\int (1 - y^{n_i} - (1-y)^{n_i} - g_2(y,n_i)) (s_i(1) - s_i(y)) \right] dm_i(y) \\
\Lambda_3 &= \frac{1}{s(1)} \int_0^1 g_2(x,n_1) g_2(x,n_2) (s(1) - s(x))\, dm(x)
\end{aligned}
\tag{3.11}
$$

Beware that $\gamma = 0$ for silent mutations although to be concise we merged the formulas of silent and replacement mutations together.

It follows from Theorem 3.1 and equation (3.10) that likelihood function of entries in a single DOHRS table can be expressed as

$$
\begin{aligned}
&L(t_1, t_2, \nu_1, \nu_2, \gamma, \theta_s, \theta_r | F_s, O_s, B_s, F_r, O_r, B_r) \\
&\quad = e^{-\theta_s \Lambda_{1,s}} \frac{(\theta_s \Lambda_{1,s})^{F_s}}{F_s!} e^{-\theta_s \Lambda_{2,s}} \frac{(\theta_s \Lambda_{1,s})^{O_s}}{O_s!} e^{-\theta_s \Lambda_{3,s}} \frac{(\theta_s \Lambda_{1,s})^{B_s}}{B_s!} \\
&\quad \cdot e^{-\theta_r \Lambda_{1,r}} \frac{(\theta_r \Lambda_{1,s})^{F_r}}{F_s!} e^{-\theta_r \Lambda_{2,r}} \frac{(\theta_r \Lambda_{2,r})^{O_r}}{O_r!} e^{-\theta_r \Lambda_{3,r}} \frac{(\theta_r \Lambda_{3,r})^{B_r}}{B_r!}
\end{aligned}
\tag{3.12}
$$

CHAPTER 4

MARKOV CHAIN MONTE CARLO IMPLEMENTATION

4.1 Hierarchical Baysian Model

Theoretically speaking, we can estimate the parameters $\boldsymbol{\alpha_s}$ and $\boldsymbol{\alpha_r}$ by maximizing the likelihood function provided at the end of Chapter 3 using a single DOHRS table, which is based on a single genetic locus. However genetic materials of two closely associated species are usually identical to each other to a considerably large degree. For instance, certain genes were found to be different by only 1.2% between humans and chimpanzees, by 1.6% between humans and gorillas and by 1.8% between gorillas and chimpanzees *. Thus not many loci have sufficient information to show statistical significance of polymorphism if analyzed individually. A more practical solution is to use a joint likelihood which involves multiple genes. Furthermore, given the complexity of the likelihood function and the dimensionality of the parameter space, it is extremely difficult to achieve analytical results. Instead, a hierarchical Baysian framework incorporated with a multilayer Markov chain Monte Carlo (MCMC) simulation scheme is implemented to obtain both joint and marginal posterior distributions. (Bustamante et al. (2003))

Given the nature of mutation rates (it must be nonnegative) and also for mathematical convenience, we choose gamma distribution as the prior distribution of θ_s and θ_r since it is the conjugate prior of a Poisson distribution. Since selection coefficients may take

*See http://www.berggorilla.org/

values from the entire real number set, it is appropriate to model them as independently and identically distributed normal random variables with unknown mean μ_γ and variance σ_γ^2. Then we pick a Gaussian hyper prior for μ_γ and an inverse-gamma hyper prior for σ_γ^2.

Unlike mutation rates and selection coefficients that vary from locus to locus, the species-specific divergence times t_1, t_2 and population size ratios ν_1, ν_2 are seen as global parameters. Based on our current knowledge, there is no known distribution which matches the form of the partial likelihood that contains t_i or ν_i, non-informative priors are considered for these global parameters. In our case, proper candidates are uniform distributions.

Detailed setup for priors and hyper priors are listed below.

$$\begin{aligned} t_i &\sim U(0, t_{\max}), \qquad i = 1, 2 \\ \nu_i &\sim U(0, \nu_{\max}), \qquad i = 1, 2 \\ \theta_s &\sim \Gamma(a_s, b_s) \\ \theta_r &\sim \Gamma(a_r, b_r) \\ \gamma &\sim \phi(\mu_\gamma, \sigma_\gamma^2) \\ \mu_\gamma &\sim \phi(\mu_0, \frac{\sigma_\gamma^2}{n_0}) \\ \sigma_\gamma^2 &\sim \Gamma^{-1}(a_0, b_0) \end{aligned} \tag{4.1}$$

where $a_s, b_s, a_r, b_r, n_0, a_0, b_0$ are constants, $\phi(x|\mu, \sigma^2) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$ is the probability density function of a Gaussian distribution, $\Gamma(x|a, b) = \frac{b^a}{\Gamma(a)} x^{a-1} e^{-bx}$ is the gamma density function and $\Gamma^{-1}(x|a, b) = \frac{b^a}{\Gamma(a)} x^{-a-1} e^{-\frac{b}{x}}$ is the inverse gamma density.

Let N_g be the total number of genetic loci included in the analysis, $n_{i,j}(i = 1, 2; 1 \leq j \leq N_g)$ be the number of DNA sequences from jth locus of species i and $h(\mu_\gamma, \sigma_\gamma, t_1, t_2, \nu_1, \nu_2, \gamma, \theta_s, \theta_r)$ denote the joint prior distribution. Under the conditions described in (4.1), the posterior distribution is proportional to the following quantity

$$
\begin{aligned}
& L(\mu_\gamma, \sigma_\gamma, t_1, t_2, \nu_1, \nu_2, \gamma, \theta_s, \theta_r | F_s, O_s, B_s, F_r, O_r, B_r) \\
& \times h(\mu_\gamma, \sigma_\gamma, t_1, t_2, \nu_1, \nu_2, \gamma, \theta_s, \theta_r) \\
& = \prod_{j=1}^{N_g} \Big[Pois_1(t_1, t_2, \nu_1, \nu_2, \theta_{s,j} | F_{s,j}, n_{1,j}, n_{2,j}) \\
& \quad \times Pois_2(t_1, t_2, \nu_1, \nu_2, \theta_{s,j} | O_{s,j}, n_{1,j}, n_{2,j}) \\
& \quad \times Pois_3(t_1, t_2, \nu_1, \nu_2, \theta_{s,j} | B_{s,j}, n_{1,j}, n_{2,j}) \\
& \quad \times Pois_1(t_1, t_2, \nu_1, \nu_2, \theta_{s,j}, \gamma_j | F_{r,j}, n_{1,j}, n_{2,j}) \\
& \quad \times Pois_2(t_1, t_2, \nu_1, \nu_2, \theta_{s,j}, \gamma_j | O_{r,j}, n_{1,j}, n_{2,j}) \\
& \quad \times Pois_3(t_1, t_2, \nu_1, \nu_2, \theta_{s,j}, \gamma_j | B_{r,j}, n_{1,j}, n_{2,j}) \\
& \quad \times \Phi(\gamma_j | \mu_\gamma, \sigma_\gamma^2) \Gamma(\theta_{s,j} | a_s, b_s) \Gamma(\theta_{r,j} | a_r, b_r) \Big] \\
& \times \Gamma^{-1}(\sigma_\gamma^2 | a_0, b_0) \phi(\mu_\gamma | \mu_0, \frac{\sigma_\gamma^2}{n_0}) \\
& \times U(t_1 | 0, t_{\max}) U(t_2 | 0, t_{\max}) U(\nu_1 | 0, \nu_{\max}) U(\nu_2 | 0, \nu_{\max})
\end{aligned}
\tag{4.2}
$$

Obtaining a full target posterior distribution over a set of parameters is essential in Baysian inference since the idea of MCMC sampling is that we estimate the expectation of a parameter by taking average over samples drawn from its corresponding posterior distribution. However computing the normalizing constant for our model and explicitly giving out the full posterior distribution is intractable due to the high dimensionality of the parameter space. Hence a multi-level MCMC sampling scheme is necessary to

handle the analysis and such effort has been made by Bustamante et al. (2003). The strategy used is Gibbs sampling, which is applicable when the joint posterior distribution involving all parameters is not known explicitly or is difficult to sample from directly, but the posterior conditional distribution of each parameter given data and other parameters is known or at least easier to sample from. Here is a brief description of the Gibbs sampling (Geman and Geman (1984)):

Algorithm 1

1. Given a known data vector $\mathbf{X}$, an unknown parameter vector $\boldsymbol{\beta} = (\beta_1, \cdots, \beta_n)$ and joint prior distribution $h(\boldsymbol{\beta})$. Sample an arbitrary initial value $\boldsymbol{\beta}^{(0)} \sim h(\boldsymbol{\beta})$.

2. For iteration $t = 1, 2, \cdots$, update parameters sequentially one by one and sweep all posterior conditional distributions, i.e.

$$\begin{aligned}
\beta_1^{(t)} &\sim p(\beta_1 | \beta_2^{(t-1)}, \cdots, \beta_n^{(t-1)}, \boldsymbol{X}) \\
&\vdots \\
\beta_j^{(t)} &\sim p(\beta_j | \beta_1^{(t)}, \cdots, \beta_{j-1}^{(t)}, \beta_{j+1}^{(t-1)}, \cdots, \beta_n^{(t-1)}, \boldsymbol{X}) \\
&\vdots \\
\beta_n^{(t)} &\sim p(\beta_n | \beta_1^{(t)}, \cdots, \beta_{n-1}^{(t)}, \boldsymbol{X})
\end{aligned} \tag{4.3}$$

   In fact, the posterior conditional distribution of β_j is proportional to the joint distribution of $\boldsymbol{\beta}$ with all other parameters taking values generated from previous iteration, that is

$$\beta_j^{(t)} \propto \pi(\beta_j) = p(\beta_1^{(t)}, \cdots, \beta_{j-1}^{(t)}, \beta_j, \beta_{j+1}^{(t-1)}, \cdots, \beta_n^{(t-1)}, \boldsymbol{X}) \tag{4.4}$$

   Furthermore, any factor in $\pi(\beta_j)$ that is not a function of β_j can be dropped out to simplify the joint distribution.

Based on the form of the joint distribution $\pi(\beta_j)$, we update each parameter by one of the following two sampling methods. First, if $\pi(\beta_j)$ matches the kernel of a well known distribution such as Gamma or Gaussian, then we can derive the corresponding posterior

conditional distribution and directly sample $\beta_j^{(t)}$ from that distribution. This sampling method is called a Gibbs-sampler. Otherwise if $\pi(\beta_j)$ does not have a closed form, we choose a jumping distribution $g(x|y)$ which is symmetric in the sense that $g(x|y) = g(y|x)$ and generate a candidate β_j' from $g(\beta_j|\beta_j^{(t-1)})$. By comparing the posterior likelihoods, accept this candidate as $\beta_j^{(t)}$ with probability $\min\{1, \frac{\pi(\beta_j')}{\pi(\beta_j^{(t)})}\}$. This updating method is called Metropolis algorithm. Next step is to examine the posterior conditional distribution of each parameter to determine the detailed updating strategies. In the following we consider parameters at a specific locus, say locus j.

The parameter γ_j

In (4.2), we pull out the three Poisson mass functions related to replacement sites and the Gaussian prior to obtain the density of γ_j,

$$\begin{aligned}\pi(\gamma_j) &= Pois_1(t_1, t_2, \nu_1, \nu_2, \theta_{s,j}, \gamma_j | F_{r,j}, n_{1,j}, n_{2,j}) \\ &\quad \times Pois_2(t_1, t_2, \nu_1, \nu_2, \theta_{s,j}, \gamma_j | O_{r,j}, n_{1,j}, n_{2,j}) \\ &\quad \times Pois_3(t_1, t_2, \nu_1, \nu_2, \theta_{s,j}, \gamma_j | F_{r,j}, n_{1,j}, n_{2,j}) \Phi(\gamma_j | \mu_\gamma, \sigma_\gamma^2) \\ &\propto \phi(\gamma_j | \mu_\gamma, \sigma_\gamma^2) \exp(-\theta_{r,j} \Lambda_{r,j}) \Lambda_{1,r,j}^{F_{r,j}} \Lambda_{2,r,j}^{O_{r,j}} \Lambda_{3,r,j}^{B_{r,j}}\end{aligned} \tag{4.5}$$

where $\Lambda = \Lambda_s(t_1, t_2, \nu_1, \nu_2, n_1, n_2) = \Lambda_{1,s} + \Lambda_{2,s} + \Lambda_{3,s}$ for silent sites, $\Lambda = \Lambda_r(t_1, t_2, \nu_1, \nu_2, \gamma, n_1, n_2) = \Lambda_{1,r} + \Lambda_{2,r} + \Lambda_{3,r}$ for replacement sites. Per the discussion above, we use Metropolis algorithm to update γ_j. The jumping distribution chosen is $U(\gamma_j - h_\gamma, \gamma_j + h_\gamma)$ with a proper random walk step size h_γ. At the tth iteration, we

1. generate a random draw γ_i' from $U(\gamma_j^{(t-1)} + h_\gamma, \gamma_j^{(t-1)} + h_\gamma)$, and then

2. generate a random value u from $U(0,1)$ and set $\gamma_j^{(t)} = \gamma_j'$ if $u \leq \dfrac{\pi(\gamma_j')}{\pi(\gamma_j^{(t-1)})}$, otherwise let $\gamma_j^{(t)} = \gamma_j^{(t-1)}$.

The parameters $\boldsymbol{\theta_{s,j}}$ and $\boldsymbol{\theta_{r,j}}$

The joint density of mutation rates $\theta_{s,j}$ and $\theta_{r,j}$ matches the kernel of Gamma distribution and hence we are able to derive their posterior conditional distributions and use the Gibbs sampler to draw updated values directly as follows

$$\begin{aligned} \theta_{s,j} &\sim \Gamma(a_s + F_{s,j} + O_{s,j} + B_{s,j}, b_s + \Lambda_s) \\ \theta_{r,j} &\sim \Gamma(a_r + F_{r,j} + O_{r,j} + B_{r,j}, b_r + \Lambda_r) \end{aligned} \tag{4.6}$$

The parameters $\boldsymbol{t_1}$ and $\boldsymbol{t_2}$

Global parameters t_1 and t_2 are involved in all Poisson mass functions across the N_g genetic loci as well as their uniform priors. The densities are given by

$$\begin{aligned} \pi(t_i) = \prod_{j=1}^{N_g} \Big[& Pois_1(t_1, t_2, \nu_1, \nu_2, \theta_{s,j} | F_{s,j}, n_{1,j}, n_{2,j}) \\ & \times Pois_2(t_1, t_2, \nu_1, \nu_2, \theta_{s,j} | O_{s,j}, n_{1,j}, n_{2,j}) \\ & \times Pois_3(t_1, t_2, \nu_1, \nu_2, \theta_{s,j} | B_{s,j}, n_{1,j}, n_{2,j}) \\ & \times Pois_1(t_1, t_2, \nu_1, \nu_2, \theta_{s,j}, \gamma_j | F_{r,j}, n_{1,j}, n_{2,j}) \\ & \times Pois_2(t_1, t_2, \nu_1, \nu_2, \theta_{s,j}, \gamma_j | O_{r,j}, n_{1,j}, n_{2,j}) \\ & \times Pois_3(t_1, t_2, \nu_1, \nu_2, \theta_{s,j}, \gamma_j | B_{r,j}, n_{1,j}, n_{2,j}) \Big] \cdot U(t_i | 0, t_{\max}) \\ \propto \prod_{j=1}^{N_g} \Big\{ & \exp\left[-(\theta_{s,j}\Lambda_{s,j} + \theta_{r,j}\Lambda_{r,j})\right] \cdot \Lambda_{1,s,j}^{F_{s,j}} \Lambda_{2,s,j}^{O_{s,j}} \Lambda_{3,s,j}^{B_{s,j}} \Lambda_{1,r,j}^{F_{r,j}} \Lambda_{2,r,j}^{O_{r,j}} \Lambda_{3,r,j}^{B_{r,j}} \Big\} \end{aligned} \tag{4.7}$$

Metropolis algorithms with uniform jumping distribution $U(t_i - h_t, t_i + h_t)$ for species $i = 1$ and 2 are applied to update the two parameters.

The parameters $\boldsymbol{\nu_1}$ and $\boldsymbol{\nu_2}$

Similarly to the case of t_1 and t_2, the population size ratios ν_1 and ν_2 appear in all Poisson terms. The corresponding densities are,

$$\begin{aligned}
\pi(\nu_i) = \prod_{j=1}^{N} \Big[& Pois_1(t_1, t_2, \nu_1, \nu_2, \theta_{s,j} | F_{s,j}, n_{1,j}, n_{2,j}) \\
& \times Pois_2(t_1, t_2, \nu_1, \nu_2, \theta_{s,j} | O_{s,j}, n_{1,j}, n_{2,j}) \\
& \times Pois_3(t_1, t_2, \nu_1, \nu_2, \theta_{s,j} | B_{s,j}, n_{1,j}, n_{2,j}) \\
& \times Pois_1(t_1, t_2, \nu_1, \nu_2, \theta_{s,j}, \gamma_j | F_{r,j}, n_{1,j}, n_{2,j}) \\
& \times Pois_2(t_1, t_2, \nu_1, \nu_2, \theta_{s,j}, \gamma_j | O_{r,j}, n_{1,j}, n_{2,j}) \\
& \times Pois_3(t_1, t_2, \nu_1, \nu_2, \theta_{s,j}, \gamma_j | B_{r,j}, n_{1,j}, n_{2,j}) \Big] \cdot U(\nu_i | 0, \nu_{\max}) \\
\propto \prod_{j=1}^{N} \Big\{ & \exp\left[-(\theta_{s,j}\Lambda_{s,j} + \theta_{r,j}\Lambda_{r,j})\right] \cdot \Lambda_{1,s,j}^{F_{s,j}} \Lambda_{2,s,j}^{O_{s,j}} \Lambda_{3,s,j}^{B_{s,j}} \Lambda_{1,r,j}^{F_{r,j}} \Lambda_{2,r,j}^{O_{r,j}} \Lambda_{3,r,j}^{B_{r,j}} \Big\}
\end{aligned} \tag{4.8}$$

Again, Metropolis algorithm is applicable to update ν_i and we still use a uniform jumping distribution of $U(\nu_i - h_\nu, \nu_i + h_\nu)$. Being ratios of population sizes, ν_1 and ν_2 are more sensitive to numerical errors. In theory any factors irrelevant to ν_i can be cancelled out when calculating the ratio $\dfrac{\pi(\nu_i')}{\pi(\nu_i^{(t-1)})}$. However computer floating point arithmetic does not do the cancellation therefore we want to remove redundant information from $\pi(\nu_i)$ as much as possible. Notice that $\Psi_{NP,1}(\boldsymbol{\alpha}, n_1)$ and $\Psi_{NP,2}(\boldsymbol{\alpha}, n_1)$ in (3.9) are related to ν_1 only, while $\Psi_{NP,1}(\boldsymbol{\alpha}, n_1)$ and $\Psi_{NP,2}(\boldsymbol{\alpha}, n_1)$ are related to ν_2 only. Let

$$\begin{aligned}
&\Psi_{LP,1}(\boldsymbol{\alpha}, n_1, n_2) = \theta\Lambda_{11},\ \Psi_{NP,1}(\boldsymbol{\alpha}, n_1) = \theta\Lambda_{12},\ \Psi_{NP,1}(\boldsymbol{\alpha}, n_2) = \theta\Lambda_{13} \\
&\Psi_{LP,2}(\boldsymbol{\alpha}, n_1, n_2) = \theta\Lambda_{21},\ \Psi_{NP,2}(\boldsymbol{\alpha}, n_1) = \theta\Lambda_{22},\ \Psi_{NP,2}(\boldsymbol{\alpha}, n_2) = \theta\Lambda_{23}
\end{aligned} \tag{4.9}$$

Thus we have $\Lambda_1 = \Lambda_{11} + \Lambda_{12} + \Lambda_{13}$ and $\Lambda_2 = \Lambda_{21} + \Lambda_{22} + \Lambda_{23}$. Then the densities of ν_1 and ν_2 are further reduced as

$$\begin{aligned}
\pi(\nu_1) \propto \prod_{j=1}^{N} \Big\{ &\exp\left[-\theta_{s,j}(\Lambda_{11,s,j} + \Lambda_{12,s,j} + \Lambda_{21,s,j} + \Lambda_{22,s,j} + \Lambda_{3,s,j})\right. \\
&\left.-\theta_{r,j}(\Lambda_{11,r,j} + \Lambda_{12,r,j} + \Lambda_{21,r,j} + \Lambda_{22,r,j} + \Lambda_{3,r,j})\right] \\
&\cdot \Lambda_{1,s,j}^{F_{s,j}} \Lambda_{2,s,j}^{O_{s,j}} \Lambda_{3,s,j}^{B_{s,j}} \Lambda_{1,r,j}^{F_{r,j}} \Lambda_{2,r,j}^{O_{r,j}} \Lambda_{3,r,j}^{B_{r,j}} \Big\} \\
\pi(\nu_2) \propto \prod_{j=1}^{N} \Big\{ &\exp\left[-\theta_{s,j}(\Lambda_{11,s,j} + \Lambda_{13,s,j} + \Lambda_{21,s,j} + \Lambda_{23,s,j} + \Lambda_{3,s,j})\right. \\
&\left.-\theta_{r,j}(\Lambda_{11,r,j} + \Lambda_{13,r,j} + \Lambda_{21,r,j} + \Lambda_{23,r,j} + \Lambda_{3,r,j})\right] \\
&\cdot \Lambda_{1,s,j}^{F_{s,j}} \Lambda_{2,s,j}^{O_{s,j}} \Lambda_{3,s,j}^{B_{s,j}} \Lambda_{1,r,j}^{F_{r,j}} \Lambda_{2,r,j}^{O_{r,j}} \Lambda_{3,r,j}^{B_{r,j}} \Big\}
\end{aligned} \tag{4.10}$$

The hyper parameters $\boldsymbol{\sigma_\gamma^2}$ and $\boldsymbol{\mu_\gamma}$

The hyper parameter σ_γ^2 has an inverse-gamma posterior conditional distribution and μ_γ is normally distributed given γ_i and σ_γ^2, that is

$$\begin{aligned}
\sigma_\gamma^2 &\sim \Gamma^{-1}\left(a_0 + \frac{N_g}{2}, b_0 + \frac{1}{2}\sum_{j=1}^{N_g}(\gamma_j - \bar{\gamma})^2 + \frac{N_g n_0(\bar{\gamma} - \mu_0)^2}{2(n_0 + N_g)}\right) \\
\mu_\gamma &\sim \phi\left(\frac{n_0\mu_0 + \sum\gamma_j}{n_0 + N_g}, \frac{\sigma_\gamma^2}{n_0 + N_g}\right)
\end{aligned} \tag{4.11}$$

4.2 Numerical Evaluation and Parallel Computing

In order to evaluate the previously derived posterior densities $\pi(\cdot)$, our major task is to calculate those complicated Λ functions numerically. A typical integral involved in the density functions can be decribed as the following form,

$$\int_0^1 \int_0^1 f(x)p(t,x,y)dm(y)dm(x) \tag{4.12}$$

If we let $u(x,t) = \int_0^1 f(x)p(t,x,y)dm(y)$, then the above integral becomes

$$\int_0^1 u(x,t)dm(x) = \int_0^1 u(x,t)\frac{e^{\gamma x}}{x(1-x)}dx \tag{4.13}$$

and it can be calculated numerically by *Gauss-Legendre(GL)* quadrature (Press et al. (1992)). The GL method approximate $\int_a^b g(x)dx$ numerically using the formula

$$\int_a^b g(x)dx = \sum_{j=1}^{n} \frac{b-a}{2} w_j f(\frac{b-a}{2}\eta_j + \frac{b+a}{2}),$$

where $\eta_j \in (-1,1),\ j = 1,\ldots,n$ are GL roots obtained by solving a polynomial equation $P_n(x) = 0$ and w_j are weights determined by

$$w_j = \frac{2}{(1-\eta_j)^2(P_n'(\eta_j))^2}, \quad j = 1,\ldots,n$$

For our case $a = 0, b = 1$ and we use $n = 20$ roots to gain the balance between accuracy of approximation and computation intensity. Thus by evaluating $u(x,t)$ at a given diffusion time t and spatial points $x_j = \frac{1}{2}\eta_j + \frac{1}{2}(j = 1,\ldots,n)$, (4.12) can be calculated as

$$\int_0^1 u(x,t)dm(x) = \sum_{j=1}^{n} \frac{1}{2} w_j u(x_j,t) \frac{e^{\gamma x_j}}{x_j(1-x_j)} \tag{4.14}$$

It has been shown that $u(x,t)$ is the solution of the following parabolic partial differential equation (PDE) with associated initial and boundary conditions,

$$\begin{aligned} &\frac{\partial}{\partial t}u(x,t) = x(1-x)\frac{\partial^2}{\partial x^2}u(x,t) + \gamma x(1-x)\frac{\partial}{\partial x}u(x,t) \\ &u(x,0) = f(x), \quad u(0,t) = u(1,t) = 0 \qquad \text{for } 0 \le x \le 1, t > 0 \end{aligned} \tag{4.15}$$

The *Crank-Nicolson (CrNi)* method, one of finite difference methods, is suitable to calculate (4.15) (See Crank and Nicolson (1947), Thomas (1995), Cebeci (2002)). The method is implicit in the sense that a system of equations needs to be solved to get values $u(x_1, s+dt), \ldots, u(x_n, s+dt)$ using the values $u(x_1,s), \ldots, u(x_n,s)$. The related system of equations is in the following form

$$\begin{bmatrix} c[1] & a[1] & 0 & \cdots & 0 \\ b[2] & c[2] & a[2] & & \vdots \\ 0 & b[3] & c[3] & \ddots & \\ \vdots & & \ddots & \ddots & a[n-1] \\ 0 & \cdots & & b[n] & c[n] \end{bmatrix} \begin{bmatrix} u(x_1, s+dt) \\ u(x_2, s+dt) \\ u(x_3, s+dt) \\ \vdots \\ u(x_n, s+dt) \end{bmatrix} = \begin{bmatrix} u(x_1, s) \\ u(x_2, s) \\ u(x_3, s) \\ \vdots \\ u(x_n, s) \end{bmatrix} \tag{4.16}$$

where arrays $a[\cdot], b[\cdot], c[\cdot]$ are coefficients depending on the scale function $s(x)$ and the speed measure $dm(x)$. The matrix equation (4.16) can be solved through tridiagonal matrix algorithm and it is a special case of Gaussian Elimination method (Datta (2010),Niyogi (2006)). Analyses have shown that in order to achieve solutions that are stable and immune to oscillations, the ratio of time step dt and spatial step $dx = x_{j+1} - x_j$

should be small enough (See Charney et al.,Zhidkov (1969), Crank and Nicolson (1947)). While using a constant jumping step dt across the time space, we evaluate $u(x,t)$ at nonuniform spatial points $(x_1, x_2, \ldots, x_n)$ due to the GL routine. Hence to be safe, we set the smallest increment dx to be $x_1 - x_0 = x_1 - 0 \approx 0.0034$ and then let $dt = dx/2$. Starting from $u(x_j, 0) = f(x_j)$, one needs to solve the equation (4.16) iteratively $\frac{t}{dt}$ times to get $u(x_j, t)$ and hence the computing speed is directly associated with the input divergence time parameter t.

The other type of integral in the new polymorphism components of Λ functions has the following form

$$\int_0^t \int_0^1 q(u, 0+, y) s(y)^2 dm(y) du \tag{4.17}$$

and it is slightly different from (4.12). Again if we let $v(x,t) = \int_0^1 q(t,x,y)s(y)^2 dm(y)$ then $v(x,t)$ is a solution of some parabolic PDE similar to (4.15) and therefore CrNi method can be applied for evaluation. For the outer layer of the integration 4.17, we apply the composite rule and compute the integral as

$$\int_0^t v(0+, u) du = dt \cdot \sum_{k=1}^{10} v(0+, t_k) \tag{4.18}$$

where $dt = \frac{t}{10}, t_k = k \cdot dt$. For the inner layer, a system of equations need to be solved once we have the boundary values $v(0+, t_k), k = 1, \ldots, 10$. In other words, for any fixed diffusion time t, we solve ten PDEs to evaluate (4.17). Consequently one can expect the

computation time required for solving (4.17) to be roughly ten times longer than the time required for solving (4.12).

An open-source integrated software facility, `R`, is chosen to implement our MCMC simulation. It not only provides a powerful environment for effective data storage, analysis and visualization and but also serve as a simple, mature and expressive programming language developed to support various statistical analyses. `R` is highly extensible through add-on packages which are fundamental units of shareable functions, data, compiled codes and documentation. There are over 5000 packages which can be easily downloaded from *Comprehensive R Achive Network (CRAN)* (Venables et al. (2002)). Other features such as extreme dynamism, name lookup with mutable environments and lazy evaluation of function arguments are also available in `R` (Wickham (2014)). Due to these features, users do not need much systematic planning before writing the code. However, the convenience comes at a cost of overwhelming optimization task to the compiler which in return limits the performance of `R` program. As for the extensibility, the other side of the story is that a lot of `R` codes are poorly written since many `R` users and contributors do not have formal training in programming or software development. Thus the performance in computing speed is usually far from satisfactory especially when iterations and recursive calls of multi-arguments functions are involved.

We initially coded our MCMC simulations using sole `R` language and a quick estimation showed that it will take over 60 years to run a million iterations (which is a typical requirement for our Markov chain to reach convergence). Hence we have to optimize the program in order to reap performance gains. Benchmark test confirms that the most time-consuming part in the simulation is to call the Λ functions repeatedly. This is antic-

ipated since solving hundreds of PDEs using R would take an extremely long time. One way to solve the issue of long iteration time is to write the Λ functions in other languages like C or C++ to improve the computing performance. This is feasible because a key aspect of the internal implementation of R is that the compiler and extension mechanism are carried through C language.

Several add-on packages have been built to integrate R with C++. The packages **rpp-bind, RAbstraction** and **RObjects** are all designed using C++ templates (Eddelbuettel (2013)) but none of them have been formally released through CRAN. The **Rserve** (Simon (2015)) package can establish a socket server to allow other programs such as C++ or Java to use R facilities without invoking R session or linking R library but it is not aimed at accelerating R program itself. The **Rcpp** (Eddelbuettel (2013)) package, which was first released in November 2008, has become the most popular language extension tool for R with over ninety CRAN packages depending on it as of November 2012. The package enables R users to access, extend and modify R objects at the C++ level. It greatly simplifies integrating C++ code with R by providing a flexible and extensible *application programming interface (API)* which supports various tasks. Typically, existing R code can be easily replaced with equivalent C++ code to achieve essentially better performance.

A recent update of **Rcpp** sets up a straightforward connection between C++ and R by utilizing a feature called "attribute" (Allaire and François. (2015)). This feature is named after an extension of C++ (see the FAQ of C++11 provided by Stroustrup). In a C++ source file, one can simply declare the attribute `[[Rcpp:export]]` in front of a function to be exported by using the name space Rcpp. Then by calling the `sourceCpp()` function in a R session, the source file can be parsed and any functions marked with the

`[[Rcpp:export]]` attribute are exported and made available to the current `R` environment just like a regular `R` function.

After rewriting the Λ functions using `C++`, the computing speed is boosted by almost 500 times comparing to the function coded in `R`. Nonetheless, it still takes several months to have our MCMC chains running enough iterations. The full likelihood function (4.2) and the updating strategies described in Section 4.1 suggest that the procedures of estimating Λ functions for any two different genetic loci do not communicate with each other due to the model assumption of independence among genes and the feature that parameters are updated one by one. For example, consider the two processes of updating the selection coefficients γ_i and γ_j for loci $i \neq j$ and one can find the following three conditions are met

- The inputs: candidates γ_i' and γ_j' are generated independently.
- The outputs: updated γ_i and γ_j based on calculated Λ function values will not replace each other.
- The input of one process does not depend on the output of the other.

The above fact is called Bernstein conditions (Bernstein (1966)) and it enables us to implement parallel algorithms in the code to increase the speed further. A classic work flow of a parallel program can be summarized as follows:

1. Set up a 'manager' process and p 'worker' processes. Initialize inputs required for the workers.

2. In an ideal case, the manager splits an incoming task into p independent subtasks and send the subtasks to the workers so that each worker gets one.

3. The manager collects outputs from workers until all p subtasks get completed.

4. Loop through step 2-3 for any subsequent tasks.

5. Shut down all processes after finishing.

In practice, we often have more subtasks than the number of workers available due to hardware restriction such as the number of CPU cores or the number of machines connected. Hence the manager will send p subtasks at the beginning of step 2 and assign the next remaining one to any worker that has completed a subtask and become idle.

Since `R` does not natively offer mechanisms to carry parallelism, a wide range of tools and packages have been invented to compensate for the demand of high-performance computing (Schmidberger et al. (2009)) [†]. Based on a message passing library specification called *Message Passing Interface (MPI)*, the package **Rmpi** pioneered the exploration by porting low level MPI functions into `R`. Whereas, the complexity and expertise required prevent a broader application of this package. The **snow** (Simple Network of Workstations) package furnishes an abstraction of lower level communication mechanisms so that a collection of workstations or a cluster of computation nodes can be used to establish the manager-worker structure. With the advent of modern technology, computers with multiple CPU cores and/or multiple CPUs have become commonplace. The trend has deep impacts on the landscape of parallel computing. The package **multicore** utilizes

[†] See also CRAN task view `https://CRAN.R-project.org/view=HighPerformanceComputing`

such multiple cores systems for parallel execution of R programs but it has limited support for machines installed with Window operation system (OS). On top of **snow** and **multicore**, the package **parallel** stands out by incorporating the functionalities of those two packages and is capable of handling large chunks of computation on various OS platforms. As a typical implementation, **parallel** offers parallelized replacements of `apply, lapply, sapply` functions that `R` users are familiar with. Moreover, for people prefer `for` loops, the **foreach** package (Weston (2015)) supplies a new looping framework that can be seen as a cross-breed of the regular `for` loops and the family of `apply` functions. It offers an operator `foreach` which automatically splits a looping task into pieces by iterators and executes those tasks in parallel if the built-in `%dopar%` operator is specified. The `foreach` loop returns a value without causing side effects as a `for` loop does and results passing from multiple workers can be combined as a list, a vector or a matrix by using `.combine` function in the package. It should be noted that **foreach** must be used in conjunction with other packages through which a backend or a cluster can be registered to enable parallel execution.

To parallelize our program, we choose a recently developed package **doParallel**, which acts as an interface connecting **foreach** and **parallel** to start a cluster and specify the number of cores used for executing tasks. Under both Unix and Window OS, a cluster can be created via a simple script in `R` like (`cl <- makeCluster(4); registerDoParallel(cl)`) in which the number 4 stands for the number of cores to be used. By using four cores on a single machine, a trial run showed that the speed of a single MCMC iteration is enhanced by three times compared to the sequentially executed program.

Specifically, we set up our parallel scheme as follows:

1. According to the number of genes N_g, create two $N_g \times 9$ matrices to store the Λ function values $\Lambda_{\cdot,s,j}$ and $\Lambda_{\cdot,r,j}$. Draw random values of parameters $(t_1, t_2, \nu_1, \nu_2, \theta_{s,j}, \theta_{r,j}, \gamma_j, \mu_\gamma, \sigma_\gamma^2)$ from priors and hyperpriors to initialize the Λ matrices.

2. Draw candidates γ_j' for $j = 1, 2, \cdots, N_g$ from the jumping distribution. Use the `foreach` loop to calculate new values of $\Lambda_{\cdot,r,j}$ in parallel. Update γ_j as well as $\Lambda_{\cdot,r,j}$ if γ_j' is accepted, otherwise return old values. Use `.combine` option in the loop to combine results as updated Λ matrix and $\boldsymbol{\gamma} = (\gamma_1, \cdots, \gamma_{N_g})$ vector so that they can be used in next step.

3. Update $\theta_{s,j}$ and $\theta_{r,j}$ via Gibbs sampler using Λ matrices from previous step.

4. Draw candidates t_1'. Calculate new Λ matrices using t_1'. Substitute the old Λ matrices with the new ones if t_1' is accepted. Update t_2 similarly.

5. Update ν_1 and ν_2 using the similar routines of updating t_1 and t_2.

6. Update μ_γ and σ_γ^2 via Gibbs sampler using the Λ matrices from the previous step.

7. Loop through step 2-6 until specified number of iterations are done.

CHAPTER 5

RESULT AND DISCUSSION

5.1 Simulation Study Under Current Model Assumptions

Initially two datasets simulated with given parameter values are used as input data to check the behavior of our model. The procedure for generating the simulated datasets is outlined as below.

1. Specify values for global parameters $(\mu_\gamma, \sigma_\gamma, t_1, t_2, \nu_1, \nu_2)$. For the first dataset, we set $\mu_\gamma = -5$ to imitate the case where the majority of replacement mutations that are observed are deleterious. As a comparison, we set $\mu_\gamma = 6$ in the second dataset. The rest of the parameters $(\sigma_\gamma, t_1, t_2, \nu_1, \nu_2)$ are set to $(3, 7, 3, 5, 3)$ and $(3.5, 8, 4, 3, 1)$ for dataset 1 and 2 respectively.

2. Generate random samples from appropriate distributions to get the number of nucleotide sequences $n_{1,j}$ and $n_{2,j}$, selection coefficient γ_j and two types of mutation rates $\theta_{s,j}$ and $\theta_{r,j}$ for jth gene ($j = 1, \ldots, N_g$, in our simulation study we set $N_g = 30$). In particular, we draw $n_{1,j}$ and $n_{2,j}$ from discrete uniform distributions on a list of consecutive integers (say 1 through 25), sample γ_j from the normal distribution with mean μ_γ and standard deviation σ_γ and generate $\theta_{s,j}$ and $\theta_{r,j}$ from continuous uniform distributions with given boundaries.

3. Generate the DOHRS Tables. At each gene, we numerically calculate means of the Poisson distributed variables F_s, O_s, B_s, F_r, O_r and B_r using the formulas (3.10)

and (3.11). Then we draw sample counts from the corresponding Poisson distributions to formulate DOHRS table entries $F_{s,j}, O_{s,j}, B_{s,j}, F_{r,j}, O_{r,j}$ and $B_{r,j}$. Eventually the simulated dataset will be a $N_g \times 8$ matrix of which each row consists of $F_{s,j}, O_{s,j}, B_{s,j}, F_{r,j}, O_{r,j}, B_{r,j}, n_{1,j}$ and $n_{2,j}$.

After a burn-in period of 200,000 MCMC iterations, we take samples every 100 step to reduce the autocorrelation. At the end of the iterations, 15,000 samples (1,500,000 iterations) were drawn for simulated dataset 1 and they were divided equally into 10 subchains. Similarly, 4,000 samples (400,000 iterations) were taken for simulated dataset 2 and formed 10 subchains with 400 samples each. To monitor the behavior of MCMC outputs and track the convergence of the chains, we generated trace plots of estimated values of each parameter and examined a variety of attributes. The recorded characteristics for each of the 10 subchains include general summary statistics such as median, mean and standard deviation as well as other diagnostic measures like Gelman-Rubin (GR) statistic (Gelman and Rubin (1992)), lower and upper bound of highest posterior density interval (HPD interval) with 95% coverage probability and autocorrelation function evaluated at lags 1, 5 and 10 respectively. These information are displayed through Table 5.1 to Table 5.4. Besides, trace plots of the global parameters $(\mu_\gamma, \sigma_\gamma, t_1, t_2, \nu_1, \nu_2)$ obtained from the 10th chain are shown in Figure 5.1 and 5.2.

For each model parameter, the GR method quantifies the convergence of multiple Monte Carlo Markov chains by comparing the within-chain variation of estimates of the parameter to the between-chain variation (see Gelman and Rubin (1992), Plummer et al. (2015)). The related diagnostic statistic is called the *potential scale reduction factor*

(PSRF) and defined based upon the ratio of the two variances described above. Any PSRF value significantly exceeding one suggests poor convergence. For simulated dataset 1, we observe stationary distributions through trace plots and most point estimates of PSRF are around 1.1 (hardly above 1.2) and hence the convergence of the chain is validated. Whereas for simulated dataset 2, the chains suffer from slow mixing issue and it is somewhat expected due to the fact we are not able to run the simulation long enough (less than 10^6 iterations). The acceptance rates of proposed γ_j and t_1 candidates are substantially lower than normal range (< 0.1) and high autocorrelations also appeared in the estimates of t_1. These facts imply that corresponding chains are trapped in low density regions.

Despite of the unsatisfactory convergence performance with regard to dataset 2, some useful inference can still be made from the information collected. For both simulated datasets, the orientation of selective effect is correctly detected. From the perspective of posterior median, we obtained precise estimates of the mean μ_γ and the standard deviation σ_γ of selection coefficient. Although the amplitudes of both parameters tend to be underestimated by the model, most 95% HPD intervals still successfully capture their true values. Another interesting discovery is that the proportion of the estimated median of the first population size ratio to that of the second population size ratio, $\frac{\hat{\nu}_1}{\hat{\nu}_2}$, is remarkably in line with the real value $\frac{\nu_1}{\nu_2}$ across all chains for both datasets. Such proportions are listed in Table 5.7 and Table 5.8. On the one hand, our model performs accurately with respect to predicting the relative ratio of the effective population size of one daughter species to that of the other. On the other hand, the current model lacks power to catch the demographic change from the ancestor to the offspring due to the

evidence that both ν_1 and ν_2 are substantially overestimated.

In addition, we are concerned with biased estimates of divergence times t_1 and t_2. Notice that simulated dataset 1 resembles dataset 2 in a manner that they both describe a situation where one daughter species experienced a longer evolution history ($\frac{t_1}{t_2} > 1$) and also expanded into a larger population size ($\frac{\nu_1}{\nu_2} > 1$). For both datasets, the results reveal a trend in underrating t_1 while overrating t_2. In order to investigate the logic behind the above bias and see whether it is simply caused by chance or by some underlying dependence between $\frac{t_1}{t_2}$ and $\frac{\nu_1}{\nu_2}$, extra simulation was run on dataset 3 with parameters $(\mu_\gamma, \sigma_\gamma, t_1, t_2, \nu_1, \nu_2)$ set to $(-5, 3, 3, 7, 5, 3)$. The related outputs are organized into similar Tables (5.5, 5.6, 5.9) and plot (Figure 5.3). The behavior of MCMC sampler mimics what we have observed with dataset 1, which should not be a surprise since dataset 3 can be treated as a revised version of dataset 1 in the sense that all parameters are the same except that t_1 and t_2 are switched. However, no significant bias in evaluating divergence times are found. This clue leads us to a second thought on the model assumptions, particularly on the way we scaled the divergence times and the population size ratios. Further discussions are presented in the next section.

Chain No.		median	mean	s.d.	G.R.	95% HPD.I		acf		
						L	U	1	5	10
1	μ_γ	-2.65	-3.27	2.29		-8.24	-0.67	0.92	0.8	0.69
	σ_γ	1.41	1.78	1.3		0.39	4.71	0.9	0.77	0.65
	t_1	6.59	5.9	2.69		0.8	9.53	0.98	0.92	0.85
	t_2	4.33	5.26	3.53		0.48	11.56	0.98	0.92	0.84
	ν_1	15.8	15.08	3.57		8.3	20	0.9	0.71	0.6
	ν_2	11.95	12.09	3.55		4.99	18.36	0.89	0.64	0.53
2	μ_γ	-3.17	-3.63	2.01	1.05	-7.52	-0.77	0.89	0.69	0.49
	σ_γ	1.7	1.97	1.13	1.05	0.4	4.19	0.84	0.64	0.46
	t_1	7.28	6.83	2.02	1.03	2.27	10	0.97	0.89	0.81
	t_2	3.49	4.02	2.56	1.05	0.4	9.74	0.96	0.87	0.8
	ν_1	15.23	14.45	4	1	7.07	20	0.93	0.78	0.68
	ν_2	12.01	11.96	3.8	1	5.32	19.42	0.9	0.7	0.55
3	μ_γ	-2.72	-3.19	1.82	1.03	-7.14	-0.71	0.87	0.62	0.44
	σ_γ	1.47	1.76	1.08	1.02	0.33	3.87	0.83	0.58	0.39
	t_1	6.26	5.92	2.36	1.02	1.24	9.71	0.98	0.9	0.81
	t_2	4.77	5.21	2.98	1.02	0.62	11.02	0.97	0.89	0.81
	ν_1	15.74	15.32	3.22	1.03	8.98	19.99	0.87	0.65	0.51
	ν_2	12.68	12.54	3.33	1.02	5.99	18.61	0.86	0.56	0.43
4	μ_γ	-3.26	-3.52	1.65	1.03	-6.61	-0.85	0.89	0.69	0.52
	σ_γ	1.76	1.94	0.99	1.03	0.44	3.9	0.83	0.64	0.48
	t_1	4.78	5.07	2.65	1.17	0.89	9.72	0.99	0.94	0.9
	t_2	6.55	6.21	3.2	1.14	0.49	10.86	0.98	0.94	0.89
	ν_1	14.54	14.12	3.89	1.02	7.15	19.95	0.95	0.83	0.74
	ν_2	12.01	11.75	3.47	1.01	4.55	17.77	0.93	0.73	0.62
5	μ_γ	-3.39	-3.8	2.02	1.03	-7.69	-0.87	0.92	0.8	0.67
	σ_γ	1.83	2.04	1.12	1.03	0.35	4.11	0.87	0.75	0.63
	t_1	6.23	5.43	3.01	1.18	0.54	9.43	0.99	0.96	0.94
	t_2	4.76	5.91	3.86	1.16	0.52	11.83	0.99	0.96	0.93
	ν_1	15.05	14.77	3.51	1.03	8.42	19.99	0.94	0.79	0.66
	ν_2	11.71	11.77	3.51	1.02	5.75	19.57	0.94	0.78	0.62

Table 5.1. Simulated Dataset 1 Summary (Chains 1-5)

Chain No.		median	mean	s.d.	G.R.	95% HPD.I		acf		
						L	U	1	5	10
6	μ_γ	-4.39	-5.18	3.27	1.22	-11.79	-0.88	0.95	0.89	0.84
	σ_γ	2.36	2.82	1.86	1.21	0.39	6.6	0.92	0.84	0.78
	t_1	2.16	2.44	1.47	1.3	0.27	5.47	0.97	0.87	0.76
	t_2	9.83	9.63	2.05	1.26	4.79	13.21	0.97	0.85	0.76
	ν_1	13.59	12.82	4.62	1.18	4.21	19.97	0.97	0.88	0.81
	ν_2	9.56	9.52	3.76	1.18	2.36	15.88	0.97	0.87	0.79
7	μ_γ	-4.47	-4.95	2.51	1.18	-9.74	-1.13	0.92	0.83	0.73
	σ_γ	2.39	2.71	1.5	1.17	0.44	5.42	0.87	0.79	0.7
	t_1	2.78	3.27	2.26	1.25	0.46	8.6	0.99	0.94	0.9
	t_2	9.31	8.88	3.15	1.22	1.2	13.73	0.98	0.93	0.87
	ν_1	11.37	11.61	4.39	1.15	5.18	19.96	0.97	0.88	0.77
	ν_2	7.98	8.51	3.84	1.16	2.55	16	0.97	0.87	0.78
8	μ_γ	-2.42	-2.78	1.48	1.21	-6.13	-0.74	0.9	0.71	0.56
	σ_γ	1.31	1.51	0.86	1.2	0.36	3.37	0.86	0.66	0.52
	t_1	4.92	4.99	2.37	1.21	0.56	8.97	0.99	0.95	0.9
	t_2	6.48	6.35	3.03	1.2	0.61	10.92	0.98	0.94	0.89
	ν_1	16.74	16.29	2.66	1.17	11.53	19.99	0.88	0.62	0.45
	ν_2	13.11	13.08	2.99	1.17	7.41	19.24	0.9	0.61	0.38
9	μ_γ	-2.85	-3.38	1.92	1.19	-7.65	-0.75	0.92	0.78	0.64
	σ_γ	1.55	1.83	1.08	1.18	0.29	4.1	0.86	0.71	0.57
	t_1	4.48	4.71	2.77	1.21	0.45	9.04	0.99	0.96	0.94
	t_2	6.87	6.69	3.65	1.2	0.66	12.43	0.99	0.96	0.94
	ν_1	14.61	14.15	3.81	1.15	7.75	19.98	0.95	0.81	0.71
	ν_2	11.29	11.18	3.91	1.14	4.39	18.14	0.96	0.82	0.72
10	μ_γ	-3.03	-3.37	1.68	1.18	-6.65	-0.78	0.9	0.74	0.6
	σ_γ	1.61	1.82	0.96	1.17	0.37	3.74	0.85	0.71	0.59
	t_1	4.28	4.15	1.78	1.2	0.82	7.25	0.98	0.9	0.79
	t_2	7.28	7.34	2.21	1.2	3.18	11.21	0.97	0.88	0.8
	ν_1	14.29	13.81	4.03	1.14	7.36	20	0.95	0.85	0.78
	ν_2	10.95	11.18	3.84	1.14	4.75	18.15	0.95	0.82	0.72

Table 5.2. Simulated Dataset 1 Summary (Chains 6-10)

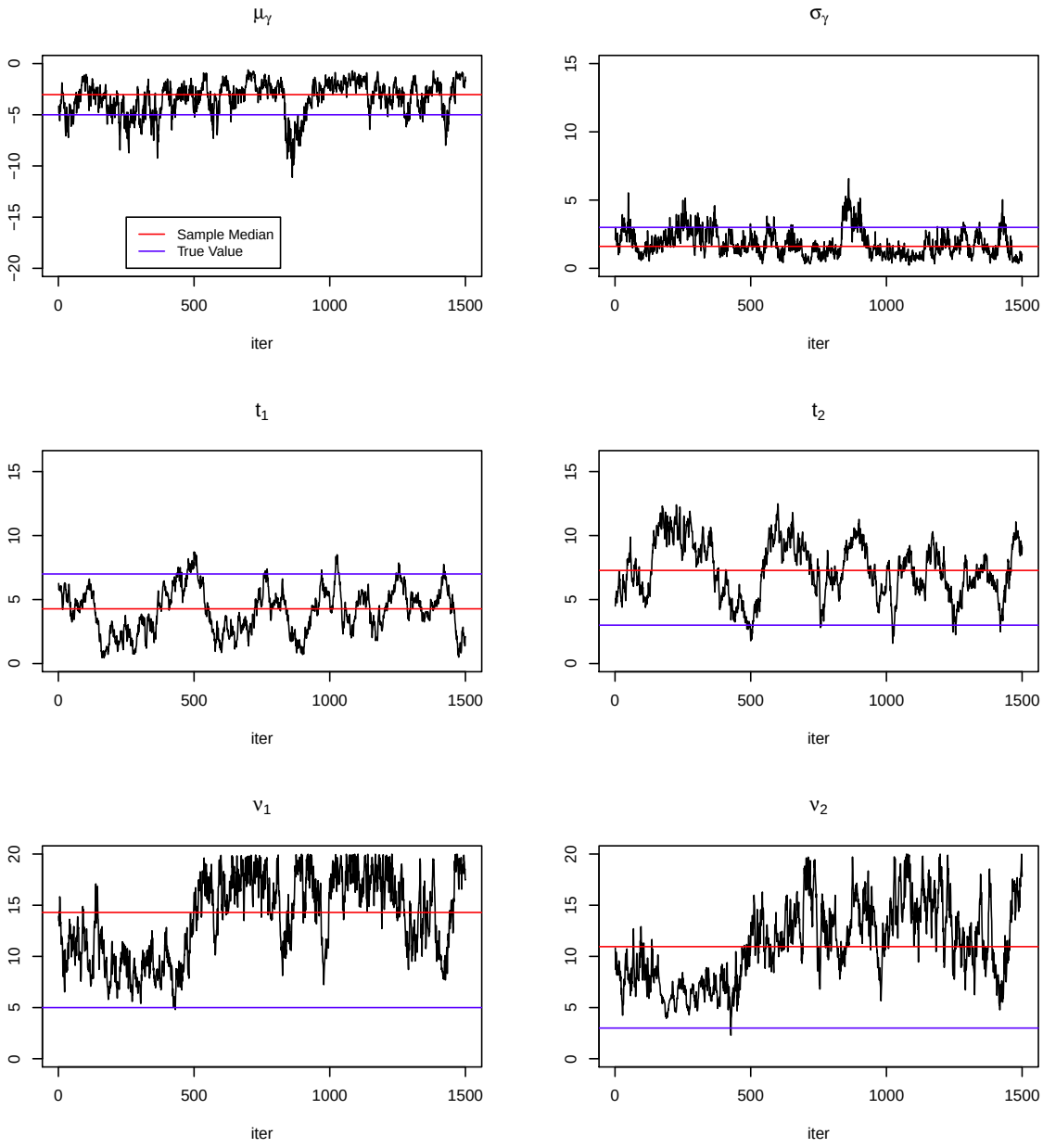

Figure 5.1. Trace Plot of Simulated Dataset 1

Chain No.		median	mean	s.d.	G.R.	95% HPD.I		acf		
						L	U	1	5	10
1	μ_γ	5.02	5.35	1.37		3.36	8.03	0.79	0.71	0.68
	σ_γ	3.19	3.29	0.92		1.73	5.02	0.69	0.64	0.57
	t_1	2.54	2.47	0.75		1.01	4.05	0.98	0.88	0.76
	t_2	18.18	19.21	3.94		13.2	27.17	0.92	0.74	0.46
	ν_1	5.27	5.25	1.3		2.96	7.57	0.97	0.88	0.75
	ν_2	1.77	1.85	0.58		0.94	2.83	0.98	0.89	0.79
2	μ_γ	4.19	4.34	0.9	1.63	2.9	6.28	0.62	0.52	0.52
	σ_γ	2.63	2.72	0.62	1.49	1.64	3.89	0.52	0.48	0.4
	t_1	3.02	3.43	1.58	1.12	1.01	6.26	0.99	0.97	0.94
	t_2	17.34	17.92	5.58	1	7.2	28.92	0.95	0.83	0.71
	ν_1	5.47	5.83	1.38	1.36	4.01	8.88	0.97	0.88	0.8
	ν_2	1.81	1.78	0.57	1.91	0.66	2.72	0.94	0.78	0.66
3	μ_γ	3.25	3.29	0.6	2.15	2.12	4.34	0.49	0.49	0.45
	σ_γ	2.03	2.11	0.46	1.79	1.36	2.96	0.45	0.38	0.39
	t_1	4.06	3.88	0.93	2.94	2.05	5.44	0.98	0.9	0.81
	t_2	19.49	21.22	7.57	1.14	10.05	36.32	0.93	0.74	0.53
	ν_1	7.11	7.47	1.52	1.37	5.36	10.29	0.98	0.92	0.87
	ν_2	1.73	1.75	0.64	1.25	0.55	2.88	0.92	0.69	0.53
4	μ_γ	2.49	2.51	0.58	2.7	1.55	3.55	0.71	0.63	0.6
	σ_γ	1.48	1.56	0.44	2.35	0.86	2.42	0.69	0.63	0.56
	t_1	3.02	3.26	1.1	2.57	1.77	5.85	0.99	0.94	0.88
	t_2	20.71	30.02	22.7	1.1	8.56	86.42	0.99	0.94	0.86
	ν_1	11.48	10.49	2.57	3.71	5.21	13.71	0.98	0.93	0.87
	ν_2	2.87	2.86	1.99	2.95	0.15	6.33	0.99	0.96	0.92
5	μ_γ	3	3.03	0.44	2.57	2.11	3.83	0.36	0.29	0.27
	σ_γ	1.82	1.85	0.34	2.23	1.26	2.48	0.25	0.15	0.09
	t_1	2.24	2.46	1.14	1.93	0.75	4.6	0.98	0.94	0.88
	t_2	19.78	20.26	4.76	1.07	10.81	30.07	0.9	0.6	0.44
	ν_1	8.91	8.92	1.16	3.21	6.43	10.95	0.94	0.75	0.52
	ν_2	3.24	3.01	0.89	2.34	0.91	4.25	0.97	0.84	0.75

Table 5.3. Simulated Dataset 2 Summary (Chains 1-5)

Chain No.		median	mean	s.d.	G.R.	95% HPD.I		acf		
						L	U	1	5	10
6	μ_γ	2.78	2.92	0.74	2.37	1.61	4.23	0.77	0.69	0.7
	σ_γ	1.87	1.95	0.53	2.05	1.14	2.97	0.65	0.61	0.57
	t_1	6.6	6.57	0.9	2.83	4.93	8.1	0.96	0.83	0.69
	t_2	7.76	8.74	5.22	1.36	1.19	19.9	0.92	0.65	0.59
	ν_1	6.29	6.29	1.39	3.05	4.19	8.77	0.98	0.92	0.85
	ν_2	2.05	2.04	0.83	2.13	0.2	3.44	0.87	0.58	0.46
7	μ_γ	4.56	4.38	1.05	2.24	2.26	6.05	0.72	0.7	0.6
	σ_γ	2.87	2.85	0.7	1.96	1.59	4.24	0.59	0.56	0.5
	t_1	5.2	5.33	1.17	2.63	3.29	7.87	0.98	0.9	0.79
	t_2	12.25	13.91	7.65	1.25	1.81	31.2	0.94	0.77	0.63
	ν_1	4.68	4.76	0.84	3.08	3.6	6.51	0.97	0.88	0.77
	ν_2	1.25	1.35	0.62	2.15	0.46	2.69	0.89	0.62	0.49
8	μ_γ	2.81	3.04	0.83	2.27	1.92	4.91	0.8	0.78	0.72
	σ_γ	1.8	1.97	0.62	1.99	1.08	3.4	0.71	0.69	0.66
	t_1	3.81	4.1	1.17	2.56	2.56	6.33	0.98	0.91	0.83
	t_2	17.9	17.91	5.94	1.24	6.75	28.82	0.9	0.75	0.59
	ν_1	8.51	7.99	2.26	3.01	4.13	11.46	0.99	0.95	0.9
	ν_2	2.13	2.18	0.88	2.03	0.7	3.77	0.94	0.79	0.66
9	μ_γ	2.09	2.16	0.58	2.42	1.21	3.26	0.73	0.7	0.59
	σ_γ	1.4	1.41	0.38	2.11	0.71	2.1	0.61	0.57	0.58
	t_1	5.35	5.82	1.49	2.6	3.7	8.64	0.98	0.91	0.84
	t_2	12.03	11.86	7.43	1.38	0.69	28.59	0.95	0.8	0.63
	ν_1	9.01	8.9	1.29	2.93	6.59	11.24	0.96	0.86	0.76
	ν_2	2.64	2.58	1.04	1.83	0.58	4.56	0.9	0.61	0.35
10	μ_γ	1.26	1.27	0.21	2.67	0.9	1.69	0.33	0.36	0.32
	σ_γ	0.8	0.82	0.17	2.34	0.55	1.15	0.34	0.31	0.28
	t_1	4.39	5.1	2.3	2.38	2.15	9.08	0.99	0.95	0.92
	t_2	12.45	10.7	5.7	1.37	0.47	18.52	0.97	0.91	0.85
	ν_1	17.17	15.68	3.69	4.46	9.67	20	0.99	0.94	0.9
	ν_2	7.13	6.78	2.23	2.93	2.13	10.71	0.96	0.82	0.65

Table 5.4. Simulated Dataset 2 Summary (Chains 6-10)

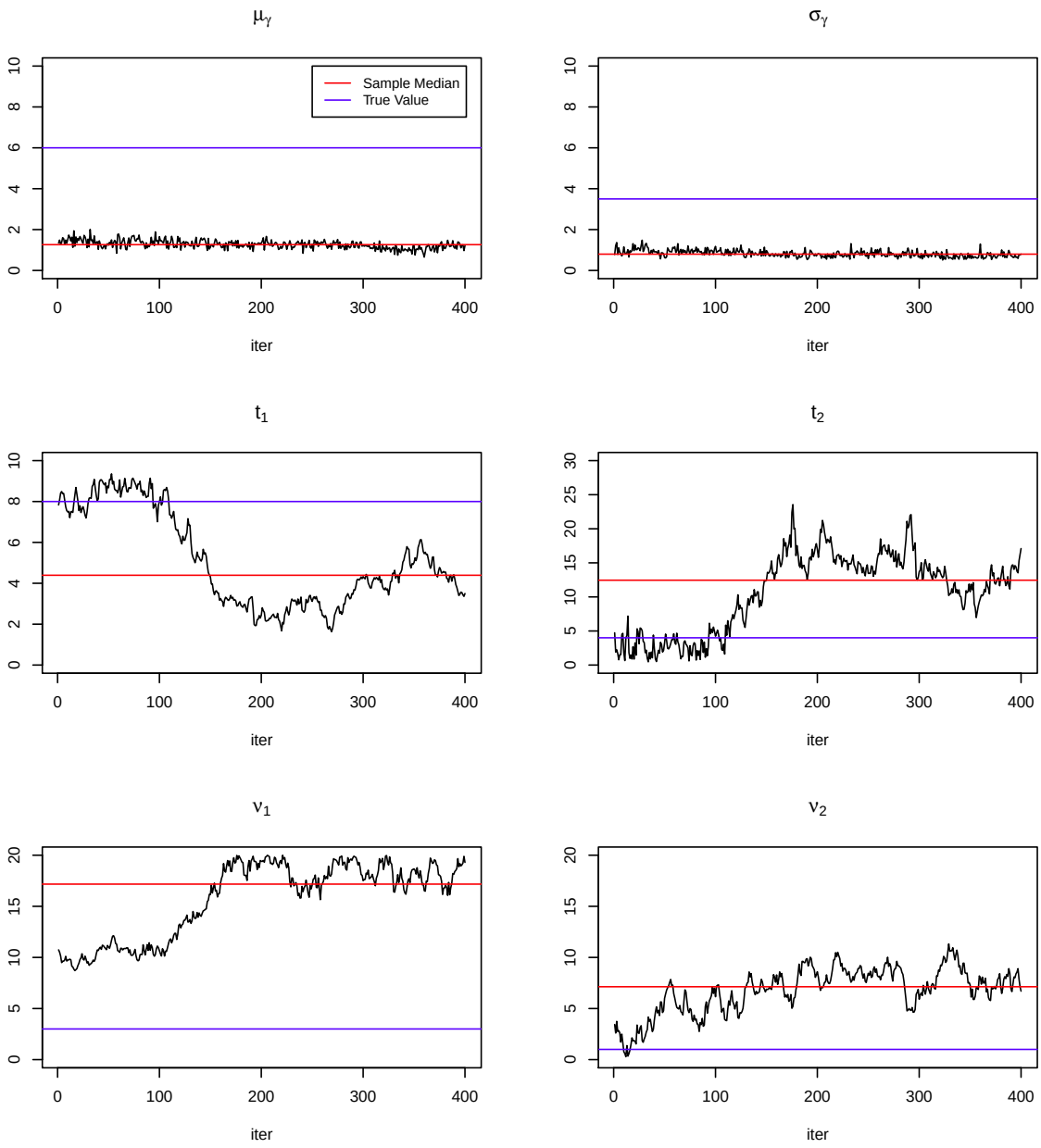

Figure 5.2. Trace Plot of Simulated Dataset 2

Chain No.		median	mean	s.d.	G.R.	95% HPD.I		acf		
						L	U	1	5	10
1	μ_γ	-2.51	-2.98	1.78		-6.73	-0.69	0.87	0.65	0.5
	σ_γ	1.44	1.76	1.15		0.3	4.18	0.83	0.6	0.47
	t_1	3.13	3.2	1.97		0.18	6.33	0.98	0.92	0.84
	t_2	7.18	6.93	3.35		0.67	12.26	0.98	0.91	0.83
	ν_1	14.1	13.71	3.46		6.13	19.22	0.99	0.97	0.95
	ν_2	7.73	7.61	2.15		3.43	11.43	0.98	0.92	0.85
2	μ_γ	-2.12	-2.58	1.53	1.62	-5.68	-0.74	0.84	0.53	0.39
	σ_γ	1.21	1.51	0.96	1.55	0.35	3.31	0.81	0.52	0.36
	t_1	2.73	2.91	1.86	1.04	0.28	6.14	0.98	0.89	0.79
	t_2	7.63	7.5	3.21	1.04	1.68	13.15	0.98	0.9	0.81
	ν_1	16.23	15.49	3.57	2.84	7.96	20	0.99	0.97	0.94
	ν_2	8.5	8.7	2.45	2.52	4.02	12.7	0.99	0.94	0.88
3	μ_γ	-2.05	-2.44	1.52	1.25	-5.68	-0.54	0.86	0.59	0.43
	σ_γ	1.17	1.44	0.97	1.22	0.27	3.46	0.81	0.54	0.38
	t_1	3.58	3.8	2.38	1.09	0.44	7.55	0.99	0.94	0.89
	t_2	6.14	5.98	3.6	1.08	0.17	10.88	0.98	0.94	0.89
	ν_1	14.64	14.37	3.27	1.71	8.68	19.72	0.99	0.97	0.94
	ν_2	9.34	9.71	2.86	1.61	4.55	15.24	0.99	0.96	0.93
4	μ_γ	-2.27	-3.11	2.47	1.19	-8.56	-0.6	0.92	0.8	0.7
	σ_γ	1.33	1.83	1.49	1.17	0.29	5.11	0.9	0.77	0.69
	t_1	2.72	3.16	1.91	1.07	0.15	6.81	0.98	0.89	0.8
	t_2	7.78	6.94	2.87	1.06	0.76	10.88	0.97	0.88	0.79
	ν_1	15	14.22	3.71	1.64	7.16	19.64	0.99	0.97	0.94
	ν_2	8.64	8.83	3.37	1.39	2.6	16.76	0.99	0.96	0.92
5	μ_γ	-2.36	-2.96	2.09	1.13	-7.42	-0.65	0.9	0.7	0.56
	σ_γ	1.36	1.77	1.31	1.11	0.28	4.47	0.86	0.64	0.5
	t_1	2.59	3.19	2.25	1.12	0.3	7.3	0.98	0.92	0.86
	t_2	7.49	6.97	3.51	1.12	0.36	11.77	0.98	0.93	0.87
	ν_1	12.78	13.5	3.3	1.44	7.96	19.44	0.99	0.96	0.93
	ν_2	8.32	8.96	3.1	1.27	4.15	16.82	0.99	0.96	0.91

Table 5.5. Simulated Dataset 3 Summary (Chains 1-5)

Chain No.		median	mean	s.d.	G.R.	95% HPD.I		acf		
						L	U	1	5	10
6	μ_γ	-2.04	-2.62	1.83	1.1	-6.64	-0.51	0.88	0.7	0.61
	σ_γ	1.17	1.54	1.13	1.09	0.25	4.04	0.86	0.67	0.58
	t_1	4.03	3.94	2.39	1.12	0.39	7.51	0.99	0.95	0.91
	t_2	5.46	6.13	4.09	1.13	0.24	12.4	0.99	0.95	0.91
	ν_1	14.48	14.98	2.31	1.35	11.61	19.75	0.98	0.92	0.85
	ν_2	9.42	9.68	2.9	1.25	4.98	15.67	0.99	0.96	0.93
7	μ_γ	-1.58	-1.97	1.22	1.13	-4.59	-0.49	0.84	0.57	0.36
	σ_γ	0.92	1.16	0.79	1.12	0.25	2.82	0.81	0.53	0.32
	t_1	3.7	3.84	2.1	1.18	0.59	7.24	0.98	0.9	0.85
	t_2	6.01	5.77	3.18	1.19	0.57	10.5	0.98	0.9	0.84
	ν_1	16.75	16.42	2.32	1.31	12.28	19.98	0.99	0.94	0.88
	ν_2	10.39	10.91	2.81	1.26	6.92	17.4	0.99	0.95	0.92
8	μ_γ	-1.84	-2.29	1.45	1.14	-5.41	-0.45	0.86	0.65	0.47
	σ_γ	1.06	1.35	0.95	1.12	0.26	3.49	0.84	0.62	0.44
	t_1	4.67	4.54	2.12	1.2	0.31	7.52	0.98	0.92	0.85
	t_2	4.54	4.72	3.03	1.22	0.21	10	0.98	0.91	0.83
	ν_1	15.56	14.62	3.67	1.25	7.33	19.93	0.99	0.97	0.95
	ν_2	9.73	10.09	2.63	1.23	5.78	15.13	0.99	0.95	0.9
9	μ_γ	-1.58	-1.92	1.14	1.15	-4.28	-0.59	0.83	0.52	0.31
	σ_γ	0.91	1.14	0.75	1.13	0.24	2.64	0.79	0.5	0.31
	t_1	3.49	3.66	2.06	1.2	0.55	7.18	0.98	0.9	0.82
	t_2	6.17	5.88	2.86	1.22	0.53	9.86	0.98	0.91	0.83
	ν_1	16.04	15.92	2.47	1.24	11.43	19.92	0.99	0.94	0.89
	ν_2	11.31	11.5	2.58	1.26	6.24	16.32	0.99	0.95	0.91
10	μ_γ	-2.21	-2.72	1.84	1.14	-6.61	-0.49	0.88	0.67	0.47
	σ_γ	1.28	1.58	1.12	1.12	0.21	3.95	0.85	0.63	0.44
	t_1	2.78	3.23	2.55	1.18	0.18	7.54	0.99	0.96	0.91
	t_2	7.7	7.06	4.04	1.19	0.24	12.77	0.99	0.95	0.91
	ν_1	15.49	15.72	2.49	1.28	11.71	19.99	0.99	0.94	0.89
	ν_2	8.58	10.07	4.26	1.3	4.05	17.5	1	0.98	0.96

Table 5.6. Simulated Dataset 3 Summary (Chains 6-10)

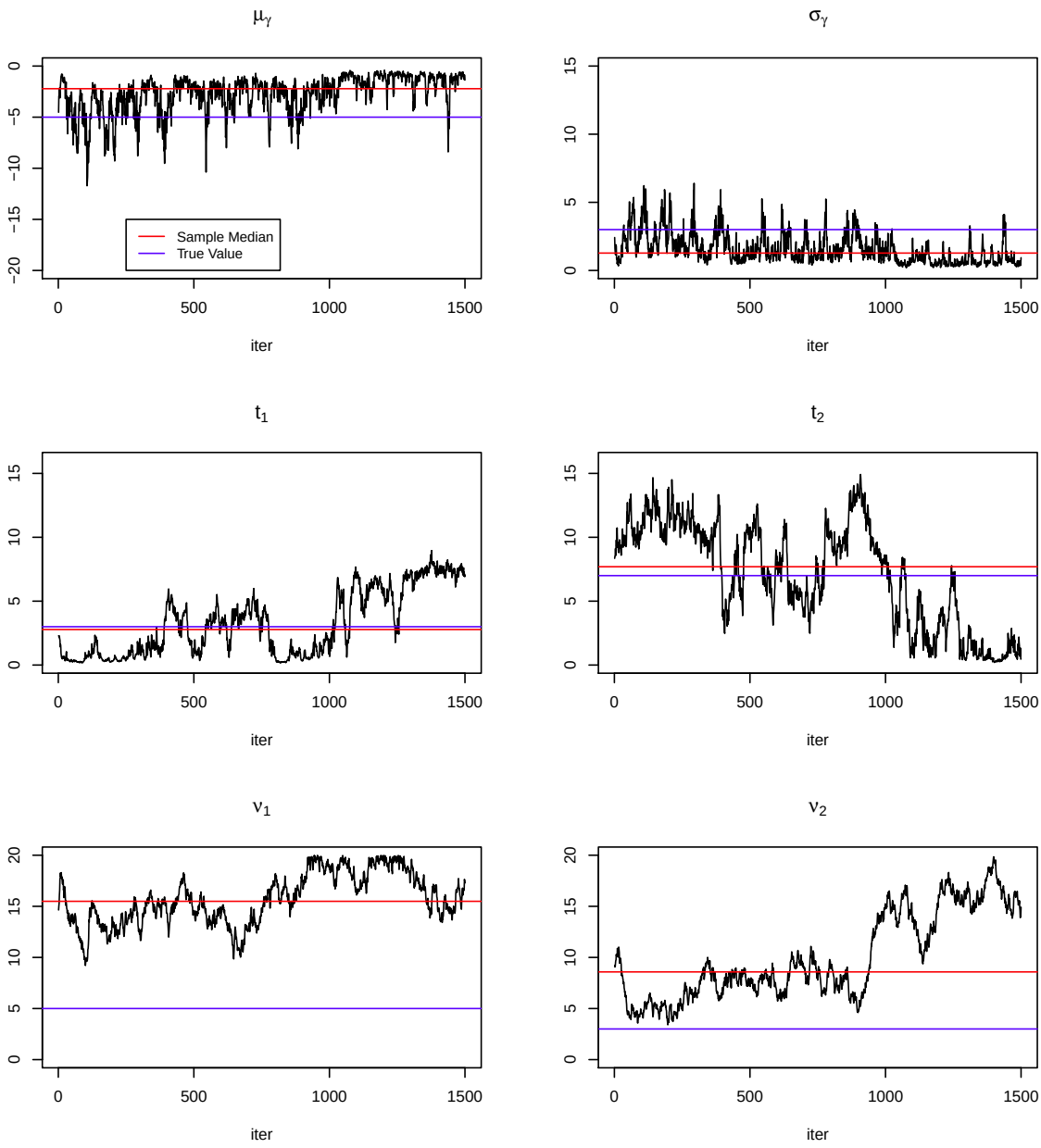

Figure 5.3. Trace Plot of Simulated Dataset 3

5.2 Further Discussion on A Model Assumption

Recall that the demographic change from the ancestor to the descendant in our model is characterized by two ratios $\nu_1 = \frac{N_1}{N_a}$ and $\nu_2 = \frac{N_2}{N_a}$, with N_1 and N_2 representing the current effective population sizes of two daughter species. In the diffusion approximation to the discrete Markov chains which describe DNA site frequency distributions, the divergence times are scaled by the haploid effective population sizes, that is

$$\frac{k_1}{N_1^2} \to t_1, \qquad \frac{k_2}{N_2^2} \to t_2$$

as $N_1, N_2 \to \infty$ respectively. In the above limits k_1 and k_2 are Moran time steps and we assume that they are independent. This setup naturally indicates that t_1 and t_2 are independent as well. However our simulation study demonstrates an evidence that the estimated ratio of divergence times $\frac{t_1}{t_2}$ tends to be less than one regardless of its true value if $\frac{\nu_1}{\nu_2}$ is set to be greater than one. Thus hidden connection may exist between the two proportions.

It may be reasonable to assume that the Moran time steps k_1 and k_2 are correlated. For mathematical simplicity, suppose that $k = k_1 = k_2$. The revised assumption yields that

$$\frac{t_1}{t_2} = \frac{k}{N_1^2} / \frac{k}{N_2^2} = (\frac{N_2}{N_1})^2 = (\frac{\nu_2}{\nu_1})^2 \tag{5.1}$$

For all three simulated datasets analyzed in Section 5.1, we compare $\frac{t_1}{t_2}$ to $(\frac{\nu_2}{\nu_1})^2$ using posterior medians. Our conjecture in (5.1) is supported by the similarity between the estimates $\frac{\hat{t}_1}{\hat{t}_2}$ and $(\frac{\hat{\nu}_2}{\hat{\nu}_1})^2$ which are listed in Table 5.7, 5.8 and 5.9. Additional analysis on datasets simulated with such revised assumption is required to reach further reliable conclusion.

Chain No.	μ_γ	σ_b	t_1	t_2	ν_1	ν_2	ν_1/ν_2	t_1/t_2	$(\frac{\nu_2}{\nu_1})^2$
1	-2.647	1.411	6.588	4.329	15.799	11.953	1.253	1.522	0.572
2	-3.168	1.701	7.275	3.493	15.227	12.007	1.205	2.083	0.622
3	-2.725	1.472	6.265	4.77	15.74	12.678	1.227	1.313	0.649
4	-3.264	1.761	4.784	6.549	14.545	12.005	1.195	0.73	0.681
5	-3.394	1.825	6.226	4.756	15.053	11.709	1.264	1.309	0.605
6	-4.389	2.356	2.158	9.826	13.588	9.559	1.356	0.22	0.495
7	-4.468	2.394	2.785	9.309	11.372	7.985	1.385	0.299	0.493
8	-2.416	1.312	4.915	6.482	16.74	13.11	1.235	0.758	0.613
9	-2.853	1.552	4.479	6.866	14.61	11.293	1.293	0.652	0.597
10	-3.03	1.608	4.277	7.279	14.293	10.951	1.247	0.588	0.587
true value	-5	3	7	3	5	3	1.667	2.333	0.36

Table 5.7. Estimated Medians for Simulated Dataset 1

Chain NO.	μ_γ	σ_b	t_1	t_2	ν_1	ν_2	ν_1/ν_2	t_1/t_2	$(\frac{\nu_2}{\nu_1})^2$
1	5.024	3.185	2.538	18.184	5.272	1.765	2.818	0.14	0.112
2	4.192	2.635	3.02	17.34	5.473	1.807	3.166	0.174	0.109
3	3.252	2.034	4.063	19.489	7.106	1.734	4.374	0.208	0.06
4	2.488	1.477	3.017	20.712	11.483	2.875	3.592	0.146	0.063
5	2.998	1.816	2.242	19.776	8.908	3.242	2.89	0.113	0.132
6	2.782	1.87	6.603	7.763	6.293	2.05	3.147	0.851	0.106
7	4.561	2.873	5.197	12.25	4.682	1.248	3.776	0.424	0.071
8	2.81	1.802	3.81	17.898	8.508	2.128	3.652	0.213	0.063
9	2.093	1.398	5.349	12.03	9.009	2.636	3.263	0.445	0.086
10	1.262	0.798	4.393	12.45	17.172	7.131	2.252	0.353	0.172
true value	6	3.5	8	4	3	1	3	2	0.111

Table 5.8. Estimated Medians for Simulated Dataset 2

Chain NO.	μ_γ	σ_b	t_1	t_2	ν_1	ν_2	ν_1/ν_2	t_1/t_2	$(\frac{\nu_2}{\nu_1})^2$
1	-2.509	1.439	3.133	7.185	14.096	7.734	1.801	0.436	0.301
2	-2.118	1.212	2.731	7.627	16.229	8.504	1.746	0.358	0.275
3	-2.045	1.17	3.578	6.14	14.645	9.344	1.536	0.583	0.407
4	-2.265	1.328	2.724	7.779	15.001	8.638	1.673	0.35	0.332
5	-2.358	1.358	2.59	7.491	12.785	8.325	1.513	0.346	0.424
6	-2.036	1.175	4.031	5.455	14.482	9.415	1.565	0.739	0.423
7	-1.58	0.917	3.701	6.014	16.749	10.389	1.551	0.615	0.385
8	-1.835	1.059	4.665	4.538	15.561	9.732	1.452	1.028	0.391
9	-1.577	0.909	3.49	6.168	16.044	11.314	1.425	0.566	0.497
10	-2.214	1.275	2.781	7.698	15.491	8.582	1.756	0.361	0.307
true value	-5	3	3	7	5	3	1.667	0.429	0.36

Table 5.9. Estimated Medians for Simulated Dataset 3

5.3 Results on The Model Application to A Drosophila Genes Data

DNA sequence data from two well-known sibling species of *Drosophila*, *D.melanogaster* and *D.simulans*, had been applied to various PRF models to make statistical inference about selection and divergence (see Bustamante et al. (2002), Pröschel et al. (2006), Sawyer et al. (2007), Amei and Sawyer (2012), Zhou (2013)). The dataset from Pröschel et al. (2006) is analyzed in our case. It contains information of 91 autosomal genes (not from sexual chromosomes) gathered from 12 *D.melanogaster* lines at Lake Kariba, Zimbabwe and multiple protein-encoding alleles are provided for each of these genes. As a interspecific comparison, genes of one *D.simulans* line from Chapel Hill, North Carolina are used to show DNA site polymorphisms. Similar to simulated datasets, FOB counts and number of alleles are organized in a 91×8 matrix with *D.melanogaster* marked as the first daughter species and *D.simulans* as the second species. It is noteworthy that the sample counts of silent and replacement legacy polymorphisms (i.e. B_s and B_r) are 0

for all genes since the sample contains only one DNA alignment from *D.simulans* and hence there is no sample polymorphism in the genes of *D.simulans*.

After dumping the initial 100,000 burn-in iterations, one MCMC draw was taken from every 50 runs and 10 subchains of equal length were created using 8,000 thinned samples. Summary reports of all global parameters are listed in Table 5.10, 5.11 and 5.12 for all 10 subchains and results from the last subchain are discussed as follows. Using the diffusion time scale presented in Chapter 3, the posterior medians together with 95% HPD credible intervals of global parameters are estimated to be 0.14 (0.04,0.29) for μ_γ, 0.27 (0.13,0.53) for σ_γ, 5.39 (2.1,6.6) for t_1, 4.37 (0, 57.07) for t_2, 13.79 (8.09, 17.49) for ν_1 and 4.24 (0,8.06) for ν_2.

The result of μ_γ indicates that the average selective effect on genes involved is nearly neutral with a slightly favorable trend. Study of the same dataset in Amei and Sawyer (2012) paper yielded a comparable median of selection coefficient of 1.98 with 95% credible interval (0.89,3.37) while Sawyer et al. (2007) results revealed a relatively strong negative selection on most non-synonymous mutations with $\mu_\gamma = -5.7$. For more details, median estimates paired with corresponding 95% credible intervals of selection coefficients for all 91 genes are increasingly sorted and visually displayed in Figure 5.4. From the perspective of directional selection, there are 70 loci whose estimated values of γ are greater than zero and 13 genes whose credible interval estimates of γ are completely over zero. In other words, only 23% genes are subject to negative selection. The overall range of the estimated selection coefficients is narrow with medians varying from -0.39 to 0.49, which is consistent with our estimates of μ_γ and σ_γ since we have assumed that γ follows a Gaussian distribution depending on those two parameters.

Using an estimated haploid effective population size of 0.645×10^6 for the ancestor species (Sawyer and Hartl (1992)), the median estimates of $t_1 = 5.39, \nu_1 = 13.79, t_2 = 4.37$ and $\nu_2 = 4.24$, suggest that *D.melanogaster* and *D.simulans* had diverged respectively 47.94 million years ago and 11.95 million years ago from their common ancestor. Meanwhile, our estimates of the population size ratios indicate that *D.melanogaster* has a larger population size compared to *D.simulans* while this result contradicts the inference made in Aquadro et al. (1988) where they hypothesized that the higher level of DNA variation observed within *D.simulans* was mainly determined by its larger population size (see also Capy and Gibert (2004)). However our estimates of ν_2 and hence the population size of *D.simulans* may be influenced by the fact that the sample size of *D.simulans* in the study is one across all genes. Consequently our result is likely to be an underestimate of the actual size.

Chain No.		median	mean	s.d.	G.R.	95% HPD.I		acf		
						L	U	1	5	10
1	μ_γ	0.1	-0.06	0.4		-1.03	0.46	0.88	0.83	0.79
	σ_γ	0.54	0.82	0.68		0.11	2.1	0.96	0.93	0.89
	t_1	2.17	3.1	1.97		0.66	6.22	0.98	0.95	0.9
	t_2	29.66	29.43	16.89		3.43	58.85	0.99	0.94	0.89
	ν_1	17.26	17.03	1.86		13.95	19.99	0.98	0.9	0.82
	ν_2	2.07	2.61	2.21		0	6.97	0.99	0.97	0.94
2	μ_γ	0.17	0.21	0.16	1.46	-0.01	0.54	0.64	0.52	0.43
	σ_γ	0.41	0.55	0.38	1.98	0.12	1.32	0.93	0.88	0.87
	t_1	3.89	3.72	1.86	1.7	0.78	6.27	0.98	0.92	0.87
	t_2	5.52	7.7	6.58	1.73	0.01	20.08	0.98	0.92	0.87
	ν_1	12.27	12.76	3.76	4.21	6.48	19.99	0.99	0.98	0.96
	ν_2	3.85	4.49	2.4	1.56	1.37	10.72	0.99	0.93	0.86
3	μ_γ	0.22	0.23	0.13	1.21	0.04	0.55	0.57	0.44	0.33
	σ_γ	0.51	0.55	0.19	1.55	0.22	0.91	0.84	0.68	0.59
	t_1	1.89	2.03	1.1	1.75	0.41	4.21	0.94	0.81	0.67
	t_2	10.53	10.36	4.57	1.59	2.06	20.22	0.97	0.9	0.82
	ν_1	14.99	14.34	3	2.11	7.59	18.73	0.99	0.94	0.88
	ν_2	5.64	5.83	1.52	1.45	2.98	8.98	0.97	0.87	0.75
4	μ_γ	0.24	0.28	0.17	1.14	0.03	0.64	0.7	0.58	0.51
	σ_γ	0.48	0.55	0.26	1.33	0.13	1.01	0.89	0.8	0.71
	t_1	2.63	2.84	1.79	1.59	0.24	5.95	0.97	0.9	0.82
	t_2	5.52	6.4	4.2	1.62	0.37	14.03	0.96	0.87	0.78
	ν_1	12	10.73	3.62	1.83	4.37	16.11	0.99	0.97	0.95
	ν_2	5.13	5.92	3.95	1.49	0.01	14.08	0.99	0.96	0.93
5	μ_γ	0.12	0.12	0.04	1.22	0.04	0.21	0.51	0.29	0.22
	σ_γ	0.22	0.22	0.06	1.49	0.11	0.33	0.76	0.55	0.5
	t_1	4.71	4.28	1.52	1.67	1.21	6.42	0.96	0.85	0.73
	t_2	1.51	2.02	1.56	1.79	0	4.8	0.94	0.81	0.7
	ν_1	14.49	14.79	2.64	1.67	10.64	19.86	0.99	0.95	0.91
	ν_2	14.03	13.62	3.08	2.21	7.85	18.97	0.98	0.92	0.84

Table 5.10. Estimation of Global Parameters (Drosophila Gene Data, Chains 1-5)

Chain No.		median	mean	s.d.	G.R.	95% HPD.I		acf		
						L	U	1	5	10
6	μ_γ	0.64	0.7	0.7	1.78	-0.62	2.51	0.85	0.77	0.69
	σ_γ	1.45	1.63	0.88	1.88	0.18	3.45	0.92	0.83	0.74
	t_1	3.9	3.48	1.84	1.68	0.54	5.91	0.97	0.91	0.84
	t_2	2.14	7.29	9.94	1.88	0	29.52	0.99	0.97	0.94
	ν_1	2.86	5.16	4.43	3.9	0.57	14.61	1	0.98	0.95
	ν_2	2.68	3.17	2.25	2.1	0.02	8.11	0.98	0.91	0.84
7	μ_γ	0.24	0.5	0.58	1.78	-0.02	1.62	0.84	0.76	0.71
	σ_γ	0.46	0.99	1.01	1.9	0.16	3.12	0.97	0.9	0.82
	t_1	3.85	3.76	1.41	1.59	1.34	6.36	0.95	0.78	0.65
	t_2	3.19	3.89	2.72	1.89	0.03	9.79	0.91	0.7	0.58
	ν_1	5.89	7.98	5.59	3	0.8	16.68	1	0.98	0.97
	ν_2	3.51	4.52	3.38	1.89	0.02	10.77	0.99	0.97	0.94
8	μ_γ	0.14	0.14	0.06	1.78	0.04	0.26	0.58	0.42	0.36
	σ_γ	0.26	0.27	0.08	1.92	0.13	0.44	0.83	0.66	0.6
	t_1	3.85	3.83	1.65	1.48	0.91	6.34	0.96	0.84	0.74
	t_2	3.66	4.64	4.08	1.9	0.02	14.97	0.95	0.77	0.59
	ν_1	13.03	13.3	2.17	2.72	9.17	17.21	0.99	0.93	0.85
	ν_2	9.82	9	4.97	1.99	0	15.49	0.99	0.97	0.95
9	μ_γ	0.11	0.12	0.05	1.78	0.03	0.23	0.64	0.44	0.38
	σ_γ	0.2	0.22	0.07	1.93	0.11	0.36	0.86	0.72	0.63
	t_1	4.05	3.82	1.64	1.41	0.88	6.32	0.96	0.82	0.69
	t_2	2.51	2.59	1.64	1.9	0	5.36	0.94	0.77	0.63
	ν_1	16.65	16.15	2.72	2.47	11.34	19.99	0.99	0.95	0.89
	ν_2	12.72	13.29	3.16	1.97	8.11	18.79	0.99	0.95	0.91
10	μ_γ	0.14	0.15	0.07	1.77	0.04	0.29	0.56	0.44	0.38
	σ_γ	0.27	0.3	0.12	1.92	0.13	0.53	0.87	0.73	0.62
	t_1	5.39	4.92	1.3	1.39	2.1	6.6	0.95	0.83	0.72
	t_2	4.37	12.94	18.69	1.57	0	57.07	0.99	0.95	0.91
	ν_1	13.79	13.17	2.48	2.32	8.09	17.49	0.99	0.93	0.87
	ν_2	4.24	3.82	2.69	1.97	0	8.06	0.99	0.95	0.9

Table 5.11. Estimation of Global Parameters (Drosophila Gene Data, Chains 6-10)

Chain NO.	μ_γ	σ_b	t_1	t_2	ν_1	ν_2	ν_1/ν_2	t_1/t_2	$(\frac{\nu_2}{\nu_1})^2$
1	0.097	0.536	2.172	29.662	17.26	2.07	8.317	0.073	0.014
2	0.168	0.413	3.886	5.517	12.27	3.853	3.043	0.704	0.099
3	0.216	0.513	1.888	10.528	14.992	5.639	2.661	0.179	0.141
4	0.239	0.482	2.635	5.524	11.996	5.13	2.287	0.477	0.183
5	0.117	0.218	4.713	1.506	14.491	14.033	1.098	3.129	0.938
6	0.642	1.455	3.897	2.138	2.86	2.676	1.379	1.823	0.875
7	0.241	0.464	3.851	3.191	5.885	3.511	1.665	1.207	0.356
8	0.138	0.258	3.852	3.656	13.027	9.817	1.339	1.054	0.568
9	0.115	0.199	4.048	2.508	16.651	12.724	1.21	1.614	0.584
10	0.138	0.268	5.393	4.368	13.785	4.239	3.335	1.235	0.095

Table 5.12. Estimated Medians for Drosophila Gene Data

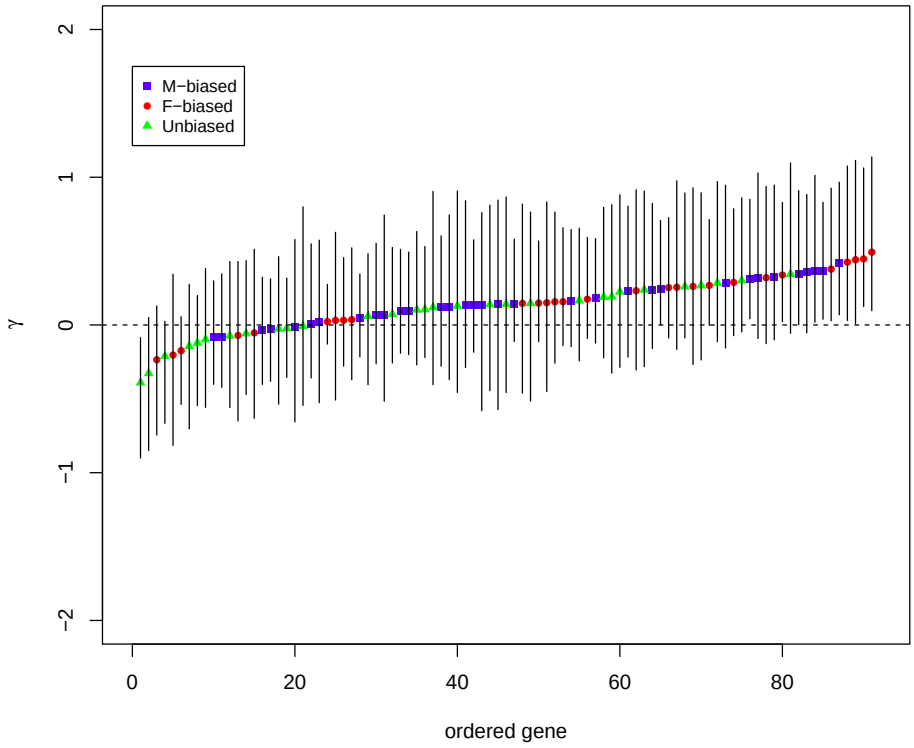

Figure 5.4. selection coefficient γ of 91 genes sorted by estimated medians

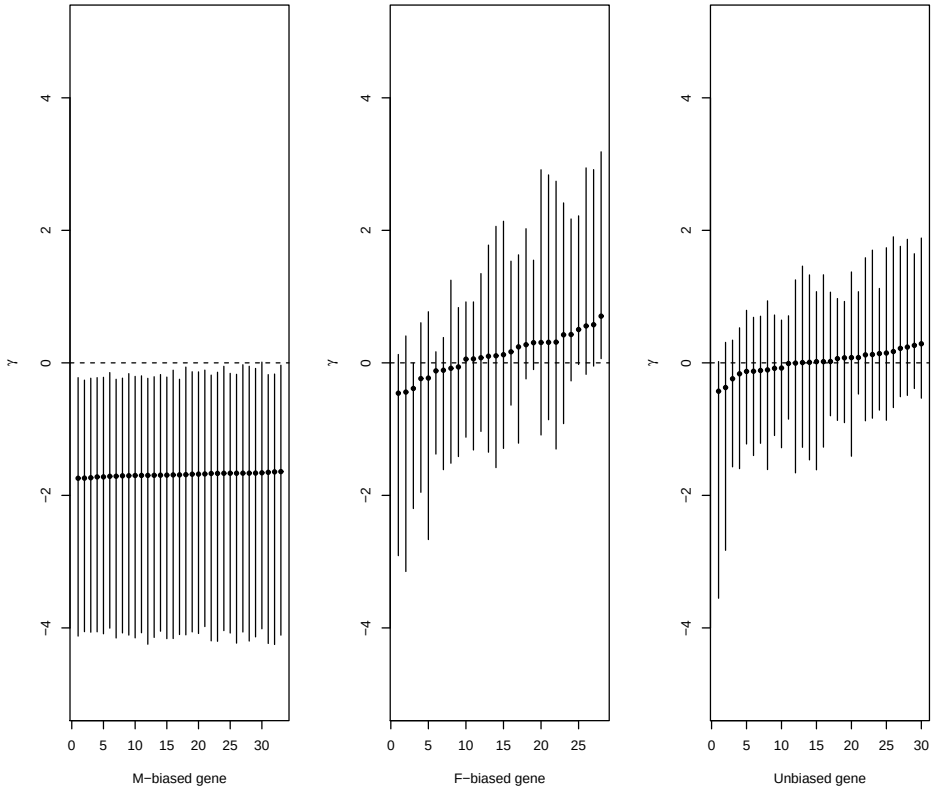

Figure 5.5. selection coefficient γ of male-biased, female-biased and unbiased subgroup

The genes in Pröschel et al. (2006) data were classified into three subgroups according to the level of genic expressions and they are groups of male-biased, female-biased and sex-unbiased genes. Analogous to Figure 5.4, estimates of selection coefficient are plotted separately in Figure 5.5 for the three subgroups. The graph suggests that sex-biased genes are more significantly subject to adaptive selection. In detail, all male-biased genes are under negative selection with a clear uniformity of intensity which matches the result obtained from Amei and Sawyer's time-dependent PRF model. Though this can be due to strong linkage among genes (Amei and Sawyer (2012)). While the selective effect on female-biased genes has a moderate variation with median estimates of scaled selection coefficients ranging from -0.46 to 0.70. In contrast, sex-unbiased genes are almost under neutral selection with a small variation between -0.42 and 0.29.

5.4 Final Remark

The classic PRF theory of Sawyer and Hartl offered a powerful approach to analyzing and interpreting intraspecific and interspecific DNA sequence polymorphism and divergence. For mathematical simplification, the original model assumes mutation-selection-drift equilibrium, leans on independence or linkage equilibrium between nucleotide sites, posits constant selection coefficient on mutations within a single genetic locus and ignores population size change over evolution history. Possible or maybe inevitable departures from these biologically unrealistic assumptions can diminish the accuracy of the model. Remarkable extensions had been made to account for such limitations. For example, Amei and Sawyer (2010) introduced a time-inhomogeneous PRF model to overcome the

issue of overestimating speciation time in time-independent models. Williamson et al. (2005) developed a population growth model using the PRF theory to infer both selection and demographic history of one species without touching interspecific comparison.

The purpose of this thesis is to relax the constant population size assumption imposed on the time-dependent PRF model. Inspired by Williamson's idea, we postulate that at the time of the divergence of two sibling species from the most recent common ancestor, each descendant species experienced a sudden population size change from the ancestral size N_a to the current size $N_i (i = 1, 2,$ respectively) and quantify such change by introducing two ratios $\nu_1 = \frac{N_1}{N_a}$ and $\nu_2 = \frac{N_2}{N_a}$. Assuming independent population sizes makes it possible to model the divergence times t_1 and t_2 for the two daughter species separately. In order to estimate the distribution of selective effects and population size ratio parameters simultaneously, we derive sample configuration formulas based on population level results and implement a multi-layer Markov chain Monte Carlo simulation scheme under a hierarchical Bayesian structure. The main barrier on the road of making reliable inference is the unbearable long time required for the convergence of the Markov chains due to massive computation on solving PDEs. Our solution is linking `C++` code to R and running the program parallel on multiple CPU cores. The developed model is validated using multiple simulated datesets and we find that the model is precise in estimating mean and variation of selection coefficients as well as predicting the relative population size alternation $\frac{\nu_1}{\nu_2}$. However both ν_1 and ν_2 tend to be overrated and it potentially leads to biased estimates of t_1 and t_2. In order to explore the unclarified dependence between divergence times and population size ratios, we propose a different way of parameterization. Additional simulation study is required to tackle this issue. Finally we present a

real data example and compare the results with those from previous studies.

Our model still preserves some restrictive conditions. For instance, the assumption of linkage equilibrium between sites may be questioned when reduced combination within genes is at present (Sawyer et al. (2007)). Allowing randomness in selective effect on nucleotide substitutions within a gene is definitely appealing though fixed effect model is robust with respect to divergence time estimate (Amei and Sawyer (2012), Zhou (2013)). Challenging the model against alternative conditions can be meaningful future attempt.

BIBLIOGRAPHY

Abel, H. J. (2009). *The role of positive selection in molecular evolution: Alternative models for within-locus selective effects*. Ph.D. thesis, Washington Univ. in St. Louis.

Adler, R. J. and Taylor, J. E. (2007). *Random fields and geometry*. Springer, NY.

Akashi, H. (1999). Inferring the fitness effects of dna mutations from polymorphism and divergence data: statistical power to detect directional selection under stationarity and free recombination. *Genetics*, 151:221–238.

Allaire, J. J., E. D. and François. (2015). Rcpp attributes.

Amei, A. and Sawyer, S. (2010). A time-dependent poisson random field model for ploymorphism within and between two related biological species. *The Anuals of Applied Probability.*, 20(5):1663–1696.

Amei, A. and Sawyer, S. A. (2012). Statistical inference of selection and divergence from a time-dependent poisson random field model. *PLoS ONE*, 7(4):e34413.

Aquadro, C. F., Lado, K. M., and Noon, W. A. (1988). The rosy region of drosophila melanogaster and drosophila simulans. i. contrasting levels of naturally occurring dna restriction map variation and divergence. *Genetics*, 119(4):875–888.

Bernstein, A. J. (1966). Analysis of programs for parallel processing. *IEEE Transactions on Electronic Computers*, EC-15.

Boyko, A. R., Williamson, S. H., Indap, A. R., Degenhardt, J. D.and Hernandez, R. D., Lohmueller, K. E., Adams, M. D., Schmidt, S., Sninsky, J. J., Sunyaev, S. R., and et al. (2008). Assessing the evolutionary impact of amino acid mutations in the human genome. *PLoS genetics*, 4(5):e1000083.

Bustamante, C. D., Nielsen, R., and Hartl, D. L. (2003). Maximum likelihood and baysian methods for estimating the distribution of selective effects among classes of mutations using dna polymorphism data. *Theory Popul. Biol.*, 63:91–103.

Bustamante, C. D., Nielsen, R., Sawyer, S. A., Olsen, K. M., Purugganan, M. D., and Hartl, D. L. (2002). The cost of inbreeding in arabidopsis. *Nature*, 416:531–534.

Bustamante, C. D., Wakeley, J., Sawyer, S., and Hartl, D. L. (2001). Directional selection and the site-frequency spectrum. *Genetics*, 159:1779–1788.

Capy, P. and Gibert, P. (2004). Drosophila melanogaster, drosophila simulans: so similar yet so different. *Genetica*, 120:5–16.

Cebeci, T. (2002). *Convective Heat Transfer*. Springer.

Charney, J. G., Fjortoft, R., and von Neumann, J. *Tellus*, (4).

Crank, J. and Nicolson, P. (1947). A practical method for numerical evaluation of solutions of partial differential equations of the heat conduction type. *Proc. Camb. Phil. Soc.*, 43:50–67.

Datta, B. N. (2010). *Numerical Linear Algebra and Applications, Second Edition*. SIAM.

Eddelbuettel, D. (2013). *Seamless R and C++ Integration with Rcpp*. Springer, NY.

Ewens, W. J. (2003). *Mathematical Population Genetics I. Theoretical Introduction*. Springer.

Gelman, A. and Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences. *Statistical Science*, 7(4):457–472.

Geman, S. and Geman, D. (1984). Stochastic relaxation, gibbs distributions, and the bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 6(6):721–741.

Griffiths, A. J. F., Wessler, S. R., Lewontin, R. C., and B., C. S. (2008). *Introduction to Genetic Analysis*. W. H. Freeman and Company, NY.

Gutenkunst, R. N., Hernandez, R. D., Williamson, S. H., and Bustamante, C. D. (2009). Inferring the joint demographic history of multiple populations from multidimensional snp frequency data. *PLoS genetics*, 5(10):e1000695.

Hartl, D. L., Moriyama, E., and Sawyer, S. A. (1994). Selection intensity for codon bias. *Genetics*, 227-234:138.

Karlin, S. and Taylor, H. M. (1981). *A second course in stochastic processes.* Academic Press, New York.

Lewontin, R. C. (1974). *The genetic basis of evolutionary change.* Columbia University Press, NY.

McDonald, J. and Kreitman, M. (1991). Adaptive protein evolution at the *adh* locus in drosophila. *Nature.*, 351:652–654.

Moran, P. A. P. (1959). The survival of a mutant gene under selection. *Journal of the Australian Mathematical Society*, 1:121–126.

Niyogi, P. (2006). *Introduction to Computational Fluid Dynamics.* Pearson Education India.

Plummer, M., Best, N., Cowles, K., and Vines, K. (2015). Package 'coda'. *R package vignette, URL http://cran.r-project.org/pub/R/web/packages/coda/coda.pdf.*

Press, W. H., Teukolsky, S. A., Vetterling, W. T., and Flannery, B. P. (1992). *Numerical Recipes in C.* Cambrideg University Press.

Pröschel, M., Zhang, Z., and DL, H. (2006). Widespread adaptive evolution of drosophila genes with sex-biased expression. *Genetics*, 174:893–900.

Raineri, D. (2001). *Introdution to Molecular Biology.* Blackwell Science, Inc.

Sawyer, S. A. (1994). Inferring selection and mutation from dna sequences: The mcdonald-kreitman test revisited. *Non-Neutral Evolution*, pages 77–87.

Sawyer, S. A. and Hartl, D. L. (1992). Population genetics of polymorphism and divergence. *Genectics*, 132:1161–1176.

Sawyer, S. A., Kulathinal, R. J., Bustamante, C. D., and Hartl, D. L. (2003). Bayesian analysis suggests that most amino acid replacements in drosophila are driven by positive selection. *Journal of Molecular Evolution*, 57:S154–S164.

Sawyer, S. A., Parsch, J., Zhang, Z., and Hartl, D. L. (2007). Prevalence of positive selection among nearly neutral amino acid replacements in drosophila. *Proc Natl Acad Sci USA.*, 104(16):6504–6510.

Schmidberger, M., Morgan, M., Eddelbuettel, D., Yu, H., Tierney, L., and Mansmann, U. (2009). State of the art in parallel computing with r. *Journal of Statistical Software*, 31(1).

Simon, U. (2015). Package 'rserve'. *R package vignette, URL https://cran.r-project.org/web/packages/Rserve/Rserve.pdf.*

Stroustrup, B. *C++11 FAQ.* stroustrup.com.

Thomas, J. W. (1995). *Numerical Partial Differential Equations: Finite Difference Methods.* Springer-Verlag:Berlin, New York.

Venables, W., Smith, D., and R.D.C, T. (2002). *An Introduciton to R.*

Weston, S. (2015). Using the **foreach** package. *doc@revolutionanalytics.com.*

Wickham, H. (2014). *Advanced R.* CRC Press, FL.

Williamson, S. H., Alon, A. F., and Bustamante, C. D. (2004). Population genetics of polymorphism and divergence for diploid selection models with arbitrary dominance. *Genetics*, 168:463–475.

Williamson, S. H., Hernandez, R., Fledel-Alon, A., Zhu, L., Nielsen, R., and Bustamante, C. D. (2005). Simultaneous inference of selection and population growth from patterns of variation in the human genome. *Proc Natl Acad Sci USA.*, 102(22):7882–7887.

Zhidkov, N. (1969). Analysis of numerical methods: E. isaacson and h. b. keller. *Zh. Vychisl. Mat. Mat. Fiz.*, 9:252–253.

Zhou, S. (2013). *Time-dependent random effect Poisson random field model for polymorphism within and between two related species.* Ph.D. dissertation, University of Nevada Las Vegas.

Zhu, L. and Bustamante, C. D. (2005). A composite-likelihood approach for detecting directional selection from dna sequence data. *Genetics*, 170:1411–1421.

CURRICULUM VITAE

Graduate College
University of Nevada, Las Vegas, USA

Jianbo Xu

Degrees:

Bachelor of Science, 2010
University of Science and Technology of China

Thesis Title:

Statistical inference of genetic forces using a Poisson random field model with non-constant population size

Thesis Examination Committee:

Chairperson, Amei Amei, Ph.D.
Committee Member, Chih-Hsiang Ho, Ph.D.
Committee Member, Malwane Ananda, Ph.D.
Graduate Faculty Representative, Guogen Shan, Ph.D.