12-1-2013

Time-Dependent Random Effect Poisson Random Field Model for Polymorphism within and Between Two Related Species

Shilei Zhou
University of Nevada, Las Vegas, zhous@unlv.nevada.edu

TIME-DEPENDENT RANDOM EFFECT POISSON RANDOM FIELD MODEL FOR POLYMORPHISM WITHIN AND BETWEEN TWO RELATED SPECIES

by

Shilei Zhou

Bachelor of Science
Xidian University, China
2007

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 2013

UNLV
UNIVERSITY OF NEVADA LAS VEGAS

THE GRADUATE COLLEGE

We recommend the dissertation prepared under our supervision by

Shilei Zhou

entitled

Time-Dependent Random Effect Poisson Random Field Model for Polymorphism within and Between Two Related Species

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., Committee Chair

Malwane Ananda, Ph.D., Committee Member

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

Evangelos Yfantis, Ph.D., Graduate College Representative

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

December 2013

ABSTRACT

TIME-DEPENDENT RANDOM EFFECT POISSON RANDOM FIELD MODEL FOR POLYMORPHISM WITHIN AND BETWEEN TWO RELATED SPECIES

by

Shilei Zhou

Dr. Amei Amei, Examination Committee Chair
Associate Professor of Mathematical Sciences
University of Nevada Las Vegas

Molecular evolution is partially driven by mutation, selection, random genetic drift, or combination of the three factors. To quantify the magnitude of these genetic forces, a previously developed time-dependent fixed effect Poisson random field model provides powerful likelihood and Bayesian estimates of mutation rate, selection coefficient, and species divergence time. The assumption of the fixed effect model that selection intensity is constant within a genetic locus but varies across genes is obviously biologically unrealistic, but it serves the original purpose of making statistical inference about selection and divergence between two related species they are individually at mutation-selection-drift inequilibrium. By relaxing the constant selection assumption, this dissertation derives a within-locus random effect model in which the selective intensity of non-synonymous mutation in a gene is treated as a

random sample from some underlying normal distribution and applies a Bayesian framework to make statistical inference about various genetic parameters. Also, a new N-ADAM-mixing Markov chain Monte Carlo sampler is created to provide better sampling strategy and fastens the convergence speed. Furthermore, to conquer the computational cost of the developed model this dissertation proposes a MPI parallel computing scheme which boosts the calculation speed by ten times.

ACKNOWLEDGEMENTS

I would like to express my sincere gratitude to my adviser, Dr. Amei Amei, for her invaluable guidance and support over the past six years. Without her extreme patient and kindness, none of this would be possible.

I thank all graduate students and staffs in Department of Mathematical Sciences who helped me go through my study and life at University of Nevada Las Vegas.

I thank Zhongren Zhang and Xiangrong Ma for their moral support at lowest point of my life.

Finally, I would like to thank my parents. They are always there for me with their unconditional love and support.

To My Parents

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

One of the central goals of population genetics is to characterize the various forces that shape pattern of genetic polymorphism within and between species. Genetic material or *DNA*, resides in *chromosomes*, is an instructional manual to produce all materials that is necessary to maintain functions of a living creature. The long twisted ladder-like string structure of DNA is built by paired *nucleotides* or *base pairs*. There are four types of nucleotides A(denine), C(ytosine), G(uanine), and T(hymine) following a rule to pair up, A with T and C with G. A certain portion of chromosome in connection with a particular trait defines a *gene* or *genetic locus*. Creature carries DNA in paired chromosomes, like higher plants or animals is called *Diploid*, while *haploid* such as bacteria has only one single chromosome. It is sometime useful to use term *Allele* as an expression of genetic type. An allele is a reflection of variation of gene and most common alleles are referred as wild types. Diploid organisms have one allele on each of their chromosomes as opposed to haploid carries only one allele. If a pair of alleles are the same, we call it *homozygote* and *heterozygote* otherwise.

Proteins are built from series of amino acids. Each protein is composed of an unique sequence of amino acids encoded from particular regions of genes. These regions of genes are sets of consecutive triplets of nucleotides called *codons*. Codons

designate all around 20 different amino acids although there are $4^3 = 64$ possible ways of composing triplets out of the four types of nucleotides. This non-one-to-one mapping gives enough room to condons to code all amino acids with redundancy. About half of amino acids are encoded by four different codons with various nucleotides in third position. Most of the remaining amino acids are encoded by two different codons. A nucleotide mutation can occur at one of the three sites in a codon position. A mutation is called *synonymous* (or *silent*) if it does not change the underlying encoded amino acid and a *non-synonymous* (or *replacement*) mutation results in a change to the amino acids. Most possible mutations at first and second sites are replacement and majority of silent mutations are at the third codon position. A DNA site is called *polymorphic* if there is more than one type of nucleotides at that site in a population and *monomorphic* otherwise.

Population genetics studies the frequencies of alleles in a population and its changes resulted from multiple factors, such as *selection*, *mutation*, *random genetic drift* or the mixture of the three factors. *Selection* or *Darwinian selection* could influence allele frequencies by effects of genes on its host and can be represented by *fitness*. A *fitness* of a gene is a measure defined by the expected number of surviving offspring in the next generation that are descendants of that gene. Positive fitnesses indicate favorable mutations while negative fitnesses mean deleterious mutations and zero are neutral. *Random genetic drift* is the result of changes of allele frequencies in a population due to *random mating*. In a Random mating system, every male-female pair in the population can act as, with equal chance, a set of parents to produce next

generation.

Statistical inference using Poisson random field (PRF) models are widely applied to sets of aligned DNA sequences to quantify various genetic forces, such as mutation, selection and divergence. Using aligned DNA sequences from closely related species McDonald and Kreitman[32] first proposed a test of neutrality of a 2×2 contingency table. The rows of the contingency table represent the number of *silent* or *replacement* nucleotide mutation sites, and columns are the number of sites that are either *fixed differences* between species or *polymorphisms* within spices. Here *fixed differences* are defined as nucleotide sites that are fixed within species but different between species and *polymorphisms* are sites that are different within one or more species. Given a genetic locus if the selection is considered to be neutral then the ratio of replacement fixed differences to replacement polymorphisms should be expected to equal the ratio of silent fixed differences to silent polymorphisms. Based on 30 aligned DNA sequences from the *alcohol dehydrogenase(Adh)* locus of three species of *Drosophila*, McDonald and Kreitman concluded that an excess amount of replacement fixed differences is a result of positive selections acting on advantageous mutations. Despite the simpleness and intuitiveness of this neutrality test, a quantitative estimate is needed to describe the strength and direction of the selective effect. Moreover, the complexity of the interaction among mutation, selection and random genetic drift is far beyond the scope of the McDonald-Kreitman test. The joint impact of such factors can be studied via a population genetics model. In 1992, Sawyer and Hartl [42] developed a Poisson random field (PRF)model. They showed

that the distribution of frequencies of mutant sites is a Poisson random field governed by silent and replacement mutation rates, selection coefficient and species divergence time. Application of the PRF model to *Adh* data, they were able to estimate all these parameters by Maximum-likelihood method[9, 42]. Bustamante et al[8]extended the original PRF model to a hierarchical Bayesian fixed effect model to estimate various genetic parameters over multiple loci with selection coefficients being assumed to be fixed within one locus, but normally distributed across all loci.

Although the Poisson random field model provides quantitative estimates of various genetic parameters, it also harbors some biologically unrealistic assumptions due to mathematical convenience. The model, among other things, assumes that nucleotide sites are independent or at *linkage equilibrium* which is equivalent to free of *recombination* [9]. Another assumption is that each species reaches mutation-selection-drift equilibrium after divergence. It also assumes that selective effect of replacement mutations within a locus is constant but varies across loci.

All these assumptions serve as original mathematical purposes of building up the PRF model and estimating genetic parameters. Recent studies have focused on relaxing such limitations and extended the original PRF model to more general settings. For example, the assumption of *linkage equilibrium* makes it inappropriate for genes exhibited strong linkages or reduced recombination[44]. Zhu and Bustamante[50] proposed a composite-likelihood ratio test that preserves independence but later adjusts for dependence among sites through coalescent simulations with recombination.The original PRF model assumes both populations have reached

mutation-selection-drift equilibrium after divergence and have same effective population size. While Sawyer et al. [44]argued that lacking of demographic factors, such as recent population growth, bottlenecks, and subdivision can undermine the model. Williamson et al. [47] proposed a time-dependent PRF model based on data from one species to infer demographic factor and natural selection by using the ratio of population sizes before and after size change as a demographic parameter. Using maximum-likelihood estimation method, they applied this model to single nucleotide polymorphism (SNP)data of 301 human genes and discovered that there was strong evidence for recent growth of human population, subject to widespread negative selections on replacement mutations. Boyko et al. [7] improved this approach to make inference of distribution of fitness on newly arising mutations with non-stationary demographic history. Application of their model to a SNP data set containing 20 European Americans and 15 African Americans yielded an ancient population expansion with African Americans and relatively recent bottleneck in European American population. The peak of estimated distribution of fitness was near neutrality while $30\% - 42\%$ of replacement mutations are moderately deleterious.

Simulation results have shown that the assumption of mutation-selection-drift equilibrium has been causing underestimate of divergence time, especially when divergence time is small [1]. Amei and Sawyer[2, 3] developed a time-dependent PRF model to explicitly implement divergence time into the model without population equilibrium assumption. They incorporated this model with Bayesian framework, applied to 91 genes of *D.melanogaster* and *D.simulans*, and used Markov chain Monte

Carlo methods to estimate selection coefficients and species divergence time. Application of the model to simulated data sets showed a strong consistency between estimates and true values.

Both time-independent and time-dependent Bayesian frameworks treat selective effects of replacement mutations within a locus as constant and such models are called "fixed effect" model. Specifically, fixed effect models assume that replacement mutations within one genetic locus are under constant selection and selection coefficient is normally distributed with fixed mean and variance across loci. This assumption is artificial, probably biologically unrealistic, and can potentially bias estimates of selection coefficients. Rather than fixed, selection coefficients should enjoy somewhat randomness within the same locus. To abandon articulateness and reveal biological reality, Sawyer et al. [44]proposed a Bayesian random effect model, in which the selection coefficient at a locus is assumed to be drawn from a normal distribution with a within-locus mean that varies among loci but with a constant within-locus variance. Results based on the application of their model to 91 genes in African populations of *D. simulans* and *D.melanogaster* data set showed that high proposition of fixation was driven by positive selection and majority of newly arisen nonsynonymous mutations are deleterious[44]. However, this random effect model is still under the time equilibrium setting.

To achieve an accurate estimate of divergence time and keep the freedom of selective effects being randomly changing, we develop a hierarchical Bayesian random effect model under time-dependent PRF framework by assuming normally distributed

within-locus selective effect. Central Limit theorem grants us a natural choice of normal distribution as underlying distribution of selective effects within locus[25, 44]. Other heavier-tailed distributions have been implemented into time-independent random effect model with an effort to eliminate model-dependence and results from heavier-tailed distribution are almost identical to those from normal distribution model[1]. The 91 genes in African populations of *D. simulans* and *D.melanogaster* data is applied to the proposed model to make statistical inference of selection and divergence and the results are compared with other studies.

We begin with a step-by-step derivation of the time-dependent PRF model and it's Bayesian implementation with random effect. Massive computational cost requires us to develop a parallel computing technique and that is discussed in Chapter 3 with great details. At last we test the newly developed method on sets of simulated data and on the real data set of 91 genes of African populations of *D. simulans* and *D.melanogaster* to infer selection and divergence.

CHAPTER 2

TIME-DEPENDENT POISSON RANDOM FIELD

2.1 Poisson Random Field

Let $\mathbf{N} = (N_1, N_2, ..., N_n)$ be n independent Poisson random variables with $E(N_i) = c_i$. Define a measure $\mu(A)$ on subset $A \subseteq X = \{0, 1, 2, ..., n\}$ by $\mu(i) = c_i$, so that $\mu(A) = \sum_{i \in A} c_i$. Now we define a random measure $N(A)$ on X by $N(i) = N_i$, so that $N(A) = \sum_{i \in A} N_i$ for $A \subseteq X$ and

$$E(N(A)) = \sum_{i \in A} E(N(i)) = \sum_{i \in A} E(N_i) = \sum_{i \in A} c_i = \mu(A). \tag{2.1}$$

Definition 1. A *Poisson random field*(PRF) is a random measure $(X, \mathcal{B}, N(A))$ on a measurable space $(X, \mathcal{B})$ with *mean measure* $(X, \mathcal{B}, \mu)$ if

$$E\left(e^{\int_X f(y)N(dy)}\right) = e^{\int_X (e^{f(y)}-1)\mu(dy)} \tag{2.2}$$

for all bounded $\mathcal{F}$-measurable functions $f(y)$ on X with $\int_X |f(y)|\mu(dy) < \infty$.

Let $\mathbf{N} = (N_1, N_2, ..., N_n)$ be as in (2.1). Suppose that there are N_i objects of some kind at state $i (1 \leq i \leq n)$. At a particular time, all the objects at state i move independently of one another to some points in a finite set Y. Assume that each of

the objects at state i moves to $y \in Y$ with a probability $\pi(i, y)$, where $\pi(i, y) \geq 0$ and $\sum_{y \in Y} \pi(i, y) = 1$ for $i \in X$.

Let M_y be the total number of objects that move to $y \in Y$, from all starting state i.

Lemma 1. *$\{M_y : y \in Y\}$ are independent Poisson random variables with means*

$$E(M_y) = \sum_{i=1}^{n} c_i \pi(i, y) \quad \textit{for} \quad y \in Y. \tag{2.3}$$

If $X = Y$, then $\pi(i, y)$ represents a Markov transition function on the set $X = \{0, 1, 2, ..., n\}$.

Suppose that, at each time $k = 0, 1, 2, ...$ a random number V_k of objects is placed at state $1 \in X$. We also assume that all $V_k, k \geq 0$ are independent and identical Poisson random variables with means $E(V_k) = v, v \geq 0$. Immediately after the k^{th} set of immigrants arrive, all objects move one step independently according to $\pi(i, y)$. In particular, one of the objects who arrived at state 1 at time k will move to state i at time $k + t$ with probability $\pi^{(t)}(1, i)$, where the $\pi^{(t)}$ is the t^{th} matrix power of $\pi(i, y)$. Define $N(s, i)$ as the number of objects at state i at time $s \geq 0$. The population is counted in the s^{th} step after the new immigrants arrive, thus $N(0, 1) = V_0$ and $N(s, 1) \geq V_s \geq 0$.

Lemma 2. *For each fixed s, $\{N(s, i)\}$ are independent Poisson random variables with means*

$$E\left(N(s, i)\right) = v \sum_{m=0}^{s} \pi^m(1, i) \quad \textit{for} \quad 0 \leq i \leq n \tag{2.4}$$

That is, at time s the expected number of mutations with population frequency i consists of mutations who entered the system at time $0 \leq m \leq s$ at state 1 (only one new mutant nucleotide occurs at each time step at a particular site) and finally reach state i.

2.2 The Moran Model

Suppose that a population of N haploid individuals has a mutant type a that initially has j_0 copies and the rest are "wild-type" A[35]. The development of the process takes place in continuous time and each individual's life time follows a negative exponential distribution with λ depending on whether the individual has type a or type A. The state of the system at any time t is defined by the number of mutant individuals. An individual who dies is immediately replaced by a new individual who is randomly chosen from the population immediately before the death. Suppose that each individual of type a has a lifetime T which follows a negative exponential distribution with *probability density function(pdf)* of $\lambda_1^{-1}e^{-\frac{t}{\lambda_1}}$, and the distribution of the lifetime of type A has *pdf* of $\lambda_2^{-1}e^{-\frac{t}{\lambda_2}}$. The *fitness* of mutant type a is denoted by $1+\sigma_N$, where $\sigma_N = \frac{\lambda_1-\lambda_2}{\lambda_2}$ is called *selection coefficient* which can be either positive or negative. Let X_k^N denote the number of individuals who carry the mutant type a in the population after k^{th} time step for $k \geq 0$. X_k^N is then a birth-and-death Markov chain on state space $S_N = \{0, 1, ..., N\}$ and has transition probabilities for $1 \leq j \leq N-1$

$$
\begin{aligned}
p_N(j, j+1) &= \frac{(1+\sigma_N)\frac{j}{N}(1-\frac{j}{N})}{1+\sigma_N(1-\frac{j}{N})} \\
p_N(j, j-1) &= \frac{\frac{j}{N}(1-\frac{j}{N})}{1+\sigma_N(1-\frac{j}{N})} \\
p_N(j, j) &= 1 - p_N(j, j+1) - p_N(j, j-1)
\end{aligned}
\tag{2.5}
$$

The states 0 and N are traps corresponding to the loss of mutant type a and the fixation of mutant type a respectively with $p_N(0,0) = p_N(N,N) = 1$. This is the so–call Moran model[35]. The original PRF model proposed by Sawyer and Hartl [42] is based on Wright-Fisher model which considers that all N individuals are replaced at each time step. In contrast, the Moran model puts one randomly chosen individual at risk to be replaced at each time step. Therefore, one generation of the Wright-Fisher model corresponds to N generations of the Moran model[2].

At time 0 we assume that there are M_0 sites which are polymorphic with mutant and wild type in population and M_j new mutations occur at time $j = 1, 2, ...$ The M_0 and M_j are independent Poisson with $E(M_j) = \mu_N \gg E(M_0)$. Let $X^N_{0,m,k}$ denote numbers of mutant nucleotides at these M_0 initial polymorphic sites at times $k \geq 0$ and is subject to $1 \leq m \leq M_0$. Also let $X^N_{1,n,j,k}$ represent numbers of mutant nucleotides at sites which are indexed by $n = 1, 2, ..., M_j$ and arise at time j, while $1 \leq j \leq k$ and $k \geq 0$, and by assumption $X_{1,n,j,j} = 1$. We consider that the model is under the infinite sites assumption that new mutations only occur at sites which have not been affected by neither new mutations nor initial mutations. We also assume that sites evolve independently and this is equivalent to saying that sites are in complete

linkage equilibrium within a locus. Under these assumptions $X^N_{0,m,k}$ and $X_{1,n,j,k}$ are independent and identical Markov chains with the same transition probabilities given by (2.5). Due to the identicalness we use X^N_k as in (2.5) to represent either $X^N_{0,m,k}$ or $X_{1,n,j,k}$, where, again, 0 and N are traps. X^N_k can be approximated by a diffusion process X_t on $(0,1)$ with time scaled as $t \sim k/N^2$ for large N(next section).

The number of sites at which there are i mutant nucleotides at time $k \geq 0$ for $1 \leq i \leq N$ is

$$N_k(i) = \#\{m : X^N_{0,m,k} = i\} + \#\{n : X^N_{1,n,j,k} = i\}, \tag{2.6}$$

where # represents cardinality of the set, $1 \leq m \leq M_0$, $1 \leq n \leq M_j$, $1 \leq j \leq k$. This definition illustrates that the number of polymorphic sites with population frequency of i/N comes from two sources. The first one is the initial polymorphisms at time $k = 0$ and the second one is new mutations that arose after $k = 0$ and already reached i/N at time $k > 0$.

The expected value of $N_k(i)$ is

$$E(N_k(i)) = \sum_{j=1}^{N-1} \omega^N_i p^k_N(j,i) + \mu_n \sum_{j=1}^{k} p^{k-j}_N(1,i). \tag{2.7}$$

where $\omega^N_i = E(N_0(i))$ and $p^k_N(j,i)$ is the k^{th} matrix power of $p_N(j,i)$. Then

$$\sum_{i=1}^{N} f(\frac{i}{N}) N_k(i) = \sum_{m=1}^{M_0} f\left(\frac{X^N_{0,m,k}}{N}\right) + \sum_{j=1}^{k}\sum_{n=1}^{M_r} f\left(\frac{X^N_{1,n,j,k}}{N}\right) \tag{2.8}$$

for any functions $f(x)$ on $[0,1]$

Here we assume that M_j and $N_0(i)$ are independent and follow Poisson distribution. An extension of Bartlett's theorem[30] proves that for each fixed time $k \geq 0$ $\{N_k(i)\}, i = 1, 2, ...$ are independent Poisson random variable and form a Poisson random field(PRF) on $\{1/N, 2/N, ..., 1\}$.

2.3 Limiting Diffusion Approximation

Two Markov chains $X_{0,m,k}$ and $X_{1,n,j,k}$ are intuitive but impractical to manage or make any inference when N gets larger. However, as $N \to \infty$ both chains can be approximated by a continuous-time continuous-state diffusion process X_t. To approach this limiting diffusion process we rescale time as N^2 steps of the discrete Markov chain so that $t \sim \frac{k}{N^2}$ and assume that the selection coefficient is scaled as $N \cdot \sigma_N \to \gamma$ as $N \to \infty$. Let $Y_k^N = \frac{X_k^N}{N}$ be the proportion of mutants at time k. We can show that for integer $i_N \in [0, N]$ such that $x_N = \frac{i_N}{N} \to x \in (0, 1)$ and any $\delta \geq 0$,

$$\begin{aligned} &\lim_{N\to\infty} N^2 E_{iN}\left(Y_1^N - x_N\right) = \gamma x(1-x) \\ &\lim_{N\to\infty} N^2 E_{iN}\left((Y_1^N - x_N)^2\right) = 2x(1-x) \\ &\lim_{N\to\infty} N^2 E_{iN}\left(|Y_1^N - x_N|^{2+\delta}\right) = 0 \end{aligned} \tag{2.9}$$

converge uniformly in x for $i_N = [Nx]$. By Taylor's theorem

$$\lim_{N\to\infty} N^2 E_{iN}\left(h(Y_1^N) - h(x_N)\right) = x(1-x)h''(x) + \gamma x(1-x)h'(x) = L_x h(x) \tag{2.10}$$

uniformly for $0 \leq x \leq 1$ for any $h \in C^2[0, 1]$.

In particular, the diffusion process X_t is generated by a following differential operator who has exit boundaries at 0 and 1 [14]

$$L_x = \frac{1}{2}a(x)\frac{d^x}{dx^2} + b(x)\frac{d}{dx} \tag{2.11}$$

where $a(x) = \gamma x(1-x)$ and $b(x) = 2x(1-x)$ are continuous functions on $x \in [0, 1]$. According to the Feller form we can rewrite the differential operator as

$$L_x = \frac{d}{m(dx)}\frac{d}{s(dx)} \tag{2.12}$$

where $s(x)($ $s(dx) = s'(x)dx)$ and $m(dx)$ are called *scale function* and *speed measure* respectively [27, 45]. From (2.8) we can write the scale and speed measure as

$$s(x) = \frac{1 - e^{-\gamma x}}{\gamma} \quad \text{and} \quad m(dx) = \frac{e^{\gamma x}}{x(1-x)}dx \tag{2.13}$$

where $s'(0) = 1$. At silent sites we set $\gamma = 0$, thus $s(x) = x$ and $m(dx) = \frac{1}{x(1-x)}dx$. Suppose that the diffusion process X_t has a smooth symmetric transition density $p(t, x, y) = p(t, y, x)$ [49] with respect to $m(dx)$ such that

$$\int_0^1 p(t, x, y)m(dy) = 1 \tag{2.14}$$

We define that

$$B_{01} = \{f \in C[0, 1] : f(0) = f(1) = 0\} \tag{2.15}$$

where $C[0,1]$ defines a collection of continuous functions on $[0,1]$. For $h(x) \in C^2(0,1) \cap B_{01}$ and any functions $f \in C[0,1]$, by Strum-Liouville theorem [11]

$$L_x h(x) = -f(x) \text{ for } 0 < x < 1 \tag{2.16}$$

has an unique solution

$$h(x) = \int_0^1 G(x,y) f(y) m(dy) \tag{2.17}$$

where $G(x,y)$ is a *Green function*, given by

$$G(x,y) = \frac{(s(1) - s(x \vee y))(s(x \wedge y) - s(0))}{s(1) - s(0)} \tag{2.18}$$

and satisfying

$$\int_0^1 \int_0^1 G(x,y)^2 m(dx) m(dy) < \infty. \tag{2.19}$$

It implies that $G(x,y)$ is a Hilbert-Schmidt kernel on $L^2(0,1)$ with respect to $m(dy)$[2, 38]. Thus there exists a complete orthonromal system of functions $\{\alpha_n(x)\}$(eigenfunctions) in $L^2(0,1) \cap C^2(0,1) \cap B_{01}$ and $\{\lambda_n\}$(eigenvalues), $0 < \lambda_1 \leq \lambda_n \uparrow \infty$ such that

$$\int_0^1 G(x,y) \alpha_n(y) m(dy) = \frac{\alpha_n(x)}{\lambda_n} \tag{2.20}$$

where we assume $\alpha_n(0) = \alpha_n(1) = 0$, and $\sum_{n=1}^{\infty} \frac{1}{\lambda_n^2} < \infty$. The equation (2.20) is equivalent to

$$L_x \alpha_n(x) = -\lambda_n \alpha_n(x) \tag{2.21}$$

By Mercer's theorem[38]

$$G(x,y) = \sum_{n=1}^{\infty} \frac{\alpha_n(x)\alpha_n(y)}{\lambda_n} \tag{2.22}$$

converges absolutely and uniformly for $0 \leq x, y \leq 1$, and

$$p(t,x,y) = \sum_{n=1}^{\infty} e^{-\lambda_n t}\alpha_n(x)\alpha_n(y) \tag{2.23}$$

converges uniformly for $0 \leq x, y \leq 1$ and $t \geq a > 0$. Hence

$$\frac{\partial}{\partial t}p(t,x,y) = L_x p(t,x,y) \quad t > 0, 0 < x, y < 1 \tag{2.24}$$

For $f(x) \in C[0,1]$ and $f(0) = f(1) = 0$ we define

$$u(x,t) = \int_0^1 p(t,x,y)f(y)m(dy) \tag{2.25}$$

and it can be solved by a parabolic partial differential equation(PDE)[2]

$$\frac{\partial}{\partial t}u(x,t) = L_x u(x,t) \quad u(x,0) = f(x). \tag{2.26}$$

Define

$$Q_t f(x) = E_x(f(X_t)) = \int_0^1 p(t,x,y)f(y)m(dy) \tag{2.27}$$

for $f \in B_{01}$. That is Q_t is a mapping from $B_{01} \to B_{01}$ and

$$|Q_t f(x)| \leq Ce^{-\lambda_1 t}||f|| \tag{2.28}$$

where $||f|| = \sup_{0\leq y\leq 1} |f(y)|$[2, 14, 27]. Then Q_t is a strongly-continuous and linear semi-group operator on B_{01}.

On Banach space B, Q_t has a *infinitesimal generator* A which is a linear operator defined by $Ah(x) = f(x)$ on domain $\mathcal{D}(A)$ [13, 27, 45]

$$\mathcal{D}(A) = \left\{ h \in B : \lim_{t\to 0} ||(1/t)(Q_t h - h) - f|| = 0 \text{ for some } f \in B \right\} \tag{2.29}$$

where $||.||$ defines the norm in the Banach space, and $\mathcal{D}(A)$ is dense in B. Referring to (2.28) $\mathcal{D}(A)$ is the range of resolvent operator

$$R_0 f(x) = \int_0^\infty Q_t f(x) dt = \int_0^1 G(x, y) f(y) m(dy) \tag{2.30}$$

hence $\mathcal{D}(A) = R_0(B_{01})$.

Define $B_C = \{f \in B_{01} | f(x) = 0 \text{ for some } a > 0, x \in [0, a] \cup [1 - a, 1]\}$, then $\mathcal{C} = R_0(B_c)$ is a *core* for A[45]. Then for all $h \in R_0(B_c)$ (2.10) holds[2]. Thus by Trotter theorem [45]:

Theorem 3. *For X_t and Y_j^N as above and $i_N = [Nx]$,*

$$\lim_{N\to\infty} E_{i_N} \left(f(Y^N_{N^2 t}) \right) = E_x(f(X_t)) = Q_t f(x) \tag{2.31}$$

uniformly for $0 \le x \le 1$ *for any* $f \in B_{01}$ *with* $f(0) = f(1) = 0$. *The convergence is also uniform for* $0 \le t \le T$ *for any* $T > 0$.

Thus the Markov chains $\{X_{0,m,k}\}$ and $\{X_{1,n,j,k}\}$ converge in distribution to the diffusion process $\{X_t\}$ with the infinitesimal generator (2.11) and scale and speed measure (2.12).

In an equilibrium case we need to find the limiting probability of fixation before the extinction of mutations. We define $h_N(i) = P_{i_N}(T_N^N < T_0^N)$, in which $T_k^N = \min\{j : X_j^N = k\}$, minimum time for Markov chain X_j^N to reach the state k starting at state i, is defined as *hitting time* of the Markov chain X_j^N and $T_a^N = \min\{t : X_t = a\}$ as the hitting time of the diffusion process $\{X_t\}$. With transition function $p_N(i,j)$ from (2.5) and the classical Gambler's Ruin problem[35] we have

Lemma 4. *Let* i_N *be integers with* $0 \le i_N \le N$ *and* $\frac{i_N}{N} \to x$ *for some* x, $0 \le x \le 1$. *Then*

$$\lim_{N\to\infty} P_{i_N}(T_N^N < T_0^N) = P_x(T_1 < T_0) = \frac{s(x) - s(0)}{s(1) - s(0)} = \frac{s(x)}{s(1)} \tag{2.32}$$

$s(x)$ *is the scale function, and if* $i_N = [Nx]$, *the convergence is uniform for* x,$0 \le x \le 1$.

A stronger version of Lemma 4 is to consider the "local limit" when $\frac{i_N}{N}$ approaches 0 or 1. That is

Lemma 5. *Let* i_N *be integers with* $1 \le i_N \le N$ *and* $x_N = \frac{i_N}{N}$. *Then*

$$\lim_{N\to\infty} \frac{P_{i_N}(T_N^N < T_0^N)}{s(x_N)/s(1)} = 1. \tag{2.33}$$

Similarly, if $0 \leq i_N \leq N-1$ *and* $x_N = \frac{i_N}{N}$, *then*

$$\lim_{N\to\infty} \frac{1 - P_{i_N}(T_N^N < T_0^N)}{(s(1) - s(x_N))/s(1)} = 1. \tag{2.34}$$

When $x_N \to 0$ (2.34) holds uniformly for $0 \leq x \leq 1$, and similarly (2.33) holds uniformly if $x_N \to 1$.

2.4 Dual Markov Chains and Time-Dependent PRF Model

At the build-up of the Moran model and diffusion process approximation, a new mutation starts at $\frac{1}{N} \to 0$, most of which will immediately die out at diffusion time scale. This causes the singularity of the diffusion process $\{X_t\}$ at $x = 0$. To avoid dealing with this singularity we work with a dual Markov chain and a dual diffusion process[29]. We define a dual Markov chain $\{\tilde{X}_k^N\}$ as $\{X_k^N | T_N^N < T_0^N\}$. That is $\{\tilde{X}_k^N\}$ is the number of mutant nucleotides at time k conditional on fixation of this mutation rather than extinction. Thus $\{\tilde{X}_k^N\}$ is a Markov chain on $\tilde{S}_N = \{1, 2, .., N\}$ which will never attains $\tilde{X}_k^N = 0$ and N is the absorbing boundary, with transition function

$$q_N(i,j) = P_i(X_1^N = j | T_N^N < T_0^N) = \frac{1}{h_N(i)} p_N(i,j) h_N(j) \tag{2.35}$$

where $h_N(i) = P_i(T_N^N < T_0^N)$ and P_i means conditional on $X_0^N = i$.

Similar to the diffusion process from last section we define $\tilde{Y}_j^N = \frac{\tilde{X}_j^N}{N}$ so that $0 < \tilde{Y}_j^N \leq 1$ and a corresponding diffusion approximation can be obtained as follows.

For integer $i_N \in \tilde{S}_N$ and any $\delta \geq 0$,

$$\begin{aligned}
\lim_{N\to\infty} N^2 \tilde{E}_{iN}\left(\tilde{Y}_1^N - x_N\right) &= \gamma x(1-x)\frac{1+e^{-\gamma x}}{1-e^{-\gamma x}} \\
\lim_{N\to\infty} N^2 \tilde{E}_{iN}\left((\tilde{Y}_1^N - x_N)^2\right) &= 2x(1-x) \\
\lim_{N\to\infty} N^2 \tilde{E}_{iN}\left(|\tilde{Y}_1^N - x_N|^{2+\delta}\right) &= 0
\end{aligned} \tag{2.36}$$

converge uniformly in x for $i_N = \min([Nx]+1)$ [2]. By Taylor's theorem we can again show that

$$\begin{aligned}
\lim_{N\to\infty} N^2 \tilde{E}_{iN}\left(h(\tilde{Y}_1^N) - h(x_N)\right) &= x(1-x)h''(x) + \gamma x(1-x)\frac{1+e^{-\gamma x}}{1-e^{-\gamma x}}h'(x) \\
&= \tilde{L}_x h(x)
\end{aligned} \tag{2.37}$$

uniformly for $0 \leq x \leq 1$ for any $h \in C^2[0,1]$.

The operator $\tilde{L}_x$ can be written in the Feller form as

$$\tilde{L}_x = \frac{d}{\tilde{m}(dx)}\frac{d}{\tilde{s}(dx)} \tag{2.38}$$

with scale and speed measure as:

$$\tilde{s}(x) = -\frac{1}{s(x)} \quad \text{and} \quad \tilde{m}(dx) = (s(x))^2 m(dx) \tag{2.39}$$

with the transition density $q(t,x,y)$.

Since $\lim_{x\to 0}\tilde{s}(x) = \infty$ and $\int_0^{1/2} |\tilde{s}(x)|\tilde{m}(dx) < \infty$ the boundary point 0 is an entrance boundary for $\tilde{X}_t$ generated by $\tilde{L}_x$ and 1 is an exit boundary [14, 27].

We define

$$B_1 = \{f \in C[0,1] : f(1) = 0\} \tag{2.40}$$

and the Green function associated with $\tilde{L}_x$ by

$$\tilde{G}(x,y) = \frac{G(x,y)}{s(x), s(y)} \tag{2.41}$$

Similar to the diffusion process X_t, we have

$$\int_0^1 \int_0^1 \tilde{G}(x,y)^2 \tilde{m}(dx)\tilde{m}(dy) = \int_0^1 \int_0^1 G(x,y)^2 m(dx)m(dy) < \infty. \tag{2.42}$$

where $\tilde{G}(x,y)$ is the kernel of Hibert–Schmidt with respect to $\tilde{m}(dx)$. Applying the same eigenfunction $\alpha_n(x)$ and eigenvalues λ_n we have

$$\int_0^1 \tilde{G}(x,y)\tilde{\alpha}_n(y)\tilde{m}(dy) = \frac{1}{s(x)} \int_0^1 G(x,y)s(y)\tilde{\alpha}_n(y)m(dy) = \frac{\tilde{\alpha}_n(x)}{\lambda_n} \tag{2.43}$$

where $\tilde{\alpha}_n(x) = \frac{\alpha_n(x)}{s(x)}$.

Hence

$$\tilde{G}(x,y) = \sum_{n=1}^{\infty} \frac{\tilde{\alpha}_n(x)\tilde{\alpha}_n(y)}{\lambda_n} = \frac{G(x,y)}{s(x)s(y)} \tag{2.44}$$

and

$$q(t,x,y) = \sum_{n=1}^{\infty} e^{-\lambda_n t}\tilde{\alpha}_n(x)\tilde{\alpha}_n(y) = \frac{p(t,x,y)}{s(x)s(y)} \tag{2.45}$$

Define

$$\tilde{Q}_t f(x) = \tilde{E}_x f(\tilde{X}_t) = \int_0^1 q(t,x,y) f(y) \tilde{m}(dy) \tag{2.46}$$

then the operator $\tilde{Q}_t$ has $|\tilde{Q}_t f(x)| \leq Ce^{-\lambda_1 t}||f||$ and forms a strongly continuous semigroup on B_1. There exists some $h(x)$ such that $h \in \mathcal{D}(\tilde{A}) = \tilde{R}_0(B_1)$ but $h \notin C^1[0,1]$ where $\tilde{R}_0 f(x) = \int_0^\infty \tilde{Q}_t f(x) dt = \int_0^1 \tilde{g}(x,y) f(y) \tilde{m}(dy)$. We define $B_C = \{f \in B_1 | f(x) = b \text{ for } x \in [0,a] \cup [1-a,1] \text{ for some } a,b > 0\}$, then (2.37)holds for $h \in \tilde{R}_0(B_c)$ [2]. Again by Trotter's theorem:

Theorem 6. *For $\tilde{X}_t$ and $\tilde{Y}_j^N$ as above and $i_N/N \to \infty$,*

$$\lim_{N\to\infty} \tilde{E}_{i_N} \left(f(\tilde{Y}^N_{N^2 t}) \right) = \tilde{E}_x(f(\tilde{X}_t)) = \tilde{Q}_t f(x) \tag{2.47}$$

for $0 \leq x \leq 1$ and any $f \in B_1$ with $f(1) = 0$. The convergence is uniform if $i_N = \min([Nx]+1, 1)$ for $0 \leq t \leq T$ and $T > 0$.

The two main results in the limiting PRF on $(0,1)$ describing the distribution of site polymorphisms and limiting expected number of mutations due to fixation are stated as follows[2].

Theorem 7. *Assume that $N\sigma_N \to \gamma$, $N\theta_N \to \theta$,$k_N \to k$ and that $N_k(i)$ defined in (2.6) satisfies $\omega_j^N = E(N_0(j))$. Then for $Q_t f(x)$*

$$\begin{aligned} &\lim_{N\to\infty} E\left(\sum_{i=1}^N f(\frac{i}{N}) N_{k_N}(i) \right) = \\ &\int_0^1 Q_t f(x) v(dx) + \theta \int_0^1 \frac{s(1)-s(x)}{s(1)-s(0)} \left(f(x) - Q_t f(x) \right) m(dx) \end{aligned} \tag{2.48}$$

for any $f \in C[0,1]$ *with* $f(0) = f(1) = 0$ *such that* $g(x) = f(x)/x$ *for* $x > 0$ *extends to a continuous function on* $[0,1]$.

Here $v(dx)$ is assumed to be a Borel measure on $(0,1)$ such that $\int_0^1 xv(dx) < \infty$ and

$$\lim_{N\to\infty} \sum_{j=1}^{N-1} g(\frac{j}{N})\frac{j}{N}\omega_j^N = \int_0^1 g(y)yv(dy) \tag{2.49}$$

for all $g \in C[0,1]$.

Mean density of limiting PRF in (2.48) is $g(t,\theta,\gamma,y)m(dy)$ such that

$$\begin{aligned} g(t,\theta,\gamma,y) = &\int_0^1 p(t,x,y)v(dx) + \\ &\theta\frac{s(1)-s(y)}{s(1)-s(0)} - \theta\int_0^1 \frac{s(1)-s(x)}{s(1)-s(0)}p(t,x,y)m(dx) \end{aligned} \tag{2.50}$$

where $v(dx)$ is an equilibrium mean density with

$$v(dx) = \theta\frac{s(1)-s(x)}{s(1)-s(0)}m(dx) = \frac{s(1)-s(x)}{x(1-x)}\frac{\theta e^{\gamma x}}{s(1)}dx \tag{2.51}$$

Theorem 8. *Under the condition of Theorem 7, the asymptotic expected number of mutant sites that have become fixed at mutant nucleotides by time t in the population*

is

$$\begin{aligned}\lim_{N\to\infty} E_{iN}(N_{k_n}(N)) &= \sum_{j=1}^{N-1} \omega_i^N p_N^k(j,N) + \mu_n \sum_{j=1}^{k} p_N^{k-j}(1,N) \\ &= \int_0^1 P_x(T_1 \le t) v(dx) + \frac{\theta}{s(1)} \int_0^t \tilde{P}_0(T_1 \le u) du \\ &= \frac{1}{s(1)} \left(\int_0^1 s(x) v(dx) - \int_0^1 \int_0^1 p(t,x,y) s(y) m(dy) v(dx) \right. \\ &\qquad \left. + \theta t - \theta \int_0^1 \int_0^1 q(u,0+,y) s(y)^2 m(dy) du \right)\end{aligned}$$

where $T_1 = \min\{t : X_t = 1\}$ is the hitting time when the diffusion process $\{X_t\}$ reached the right exit boundary[2].

For all three equations (2.48), (2.50) and (2.52), the first terms on the right-hand sides are due to initial polymorphic at time $t = 0$ and the rests are subject to new mutations after time $t > 0$.

CHAPTER 3

TIME-DEPENDENT RANDOM EFFECT PRF MODEL

3.1 DOHRS Table

Suppose that we have random samples from two closely related species. For example, $m+n$ sets of aligned DNA sequences are acquired from two species. Without loss of generality, we assume that m sets are from the first species and the second species contributes n sets. The infinite sites assumption guarantees that mutations are so rare that it can only occur once at a nucleotide site. We also assume that there is one of two types of alleles at each nucleotide sites, mutant type or wild type.

McDonald and Kreitman first proposed a 2×2 contingency table, also called DPRS table to describe numbers of fixed differences and polymorphisms at silent or replacement sites in a joint alignment[32]. Specifically in the following table

	Fixed Differences (D)	Polymorphisms (P)
Silent (S)	K_s	V_s
Replacement (R)	K_r	V_r

Table 3.1: DPRS Table

K_s and K_r represent numbers of silent and replacement fixed differences between two species, and V_s and V_r are silent and replacement polymorphisms within one or two species. Here *fixed differences* are defined as sites that have same allele types within each species but different between species and *polymorphisms* are sites that are polymorphic within one or two species.

In the time-dependent PRF model, an ancestral species with population size N is assumed to diverge into two species at a time point in the past and the two daughter populations have the same size N. The polymorphisms within one species can possibly come from either *legacy polymorphism* or *new mutation polymorphism*. Here *new mutation polymorphism* comes from nucleotide mutation that occurs after the divergence and *legacy polymorphism* is caused by the initial polymorphisms that already exist in the ancestor population at the time of divergence. New mutation polymorphism is assumed to be observable only within one daughter species while legacy polymorphism may be shared between two species. Therefore polymorphisms appear in both daughter populations are mainly due to legacy polymorphisms. Under this setting, the 2×2 DPRS table is extended to 2×3 contingency table, called DOHRS table, which divides the polymorphism into two columns accordingly. Similar to previous notations, K_s and K_r represent total number of silent and replacement fixed differences between two species, O_s and O_r are numbers of sites that are polymorphic in only one species, and H_s and H_r are numbers of sites that are polymorphic in both species

	Fixed Differences (D)	New Polymorphism (O)	Legacy Polymorphisms (H)
Silent(S)	K_s	O_s	H_s
Replacement(R)	K_r	O_r	H_r

Table 3.2: DOHRS Table

3.2 Sample Configuration Formulas

The six counts (K_s O_s H_s K_r O_r H_r) in the DOHRS table are independent Poisson random variables and their means depend on a set of parameters $\beta(t_{div}, \theta, \gamma)$. Assuming that both species have the same effect population size N, the t_{div} is the scaled diffusion time since the divergence of the species. In other words, the ancestor of these two species diverged $t_{div}N^2$ generations ago. We also assume that the two daughter population share same mutation rate θ and selection coefficient γ, with θ_s representing silent mutation rate and θ_r for replacement mutation rate. In the previously developed time-dependent PRF model, all replacement mutations happened at nucleotide sites within a genetic locus are assumed to have a constant selection coefficient γ and such γ varies from locus to locus. In this dissertation, we removed this biologically unrealistic assumption by treating the selective effect of nonsynonymous mutations in a gene as a random sample from some underlying distribution. The detail of the model is presented in the next section, Given that PRF models use aligned DNA sequences from two species as input data, the theoretical results presented in Chapter 2 need to be incorporated into corresponding sample configuration formulas. In a sample fixed difference sites are due to the fixation of a mutant type at a given locus or draws from polymorphic sites that by chance form a set of monomorphic

nucleotides. Sites that are polymorphic only in one species are sampled from either legacy polymorphic sites or new mutation sites and polymorphic sites shared by both species are outcomes of legacy polymorphisms.

At a given legacy polymorphic site, let x be the population frequency of a mutant nucleotide at time $t = 0$. In a random sample of n sequenced genes from a single daughter population let $I(x, n)$ to be the probability that the site is monomorphic in the sample for the wild type(nonmutant type), $J(x, n)$ be the probability that the site is polymorphic in the sample and $K(x, n)$ the probability that site is monomorphic in the sample for the mutant nucleotide. Then these probabilities are given by:

$$\begin{aligned} I(x,n) &= P_x(T_0 \le t) + \int_0^1 p(t,x,y)(1-y)^n m(dy) \\ &= 1 - \frac{s(x)}{s(1)} - \int_0^1 p(t,x,y)\left(1-(1-y)^n - \frac{s(y)}{s(1)}\right) m(dy) \end{aligned}$$

$$\begin{aligned} K(x,n) &= P_x(T_1 \le t) + \int_0^1 p(t,x,y) y^n m(dy) \\ &= \frac{s(x)}{s(1)} - \int_0^1 p(t,x,y)\left(y^n - \frac{s(y)}{s(1)}\right) m(dy) \end{aligned} \tag{3.1}$$

$$J(x,n) = \int_0^1 p(t,x,y)\left(1 - y^n - -(1-y)^n\right) m(dy)$$

At a time $t > 0$, let C_1 be the number of legacy polymorphic sites that are fixed differences in the sample, C_2 is the number of legacy polymorphic sites that are polymorphic in only one species and C_3 is the number of legacy polymorphic sites that are polymorphic in both samples and specifically, they are

$$
\begin{aligned}
C_1(\beta) &= \int_0^1 \left(I(x,m)K(x,n) + I(x,n)K(x,m)\right) v(dx) \\
C_2(\beta) &= \int_0^1 \left(J(x,m)(K(x,n)+I(x,n)) + J(x,n)(K(x,m)+I(x,m))\right) v(dx) \\
&= \int_0^1 \left(J(x,m) + J(x,n) - 2J(x,m)J(x,n)\right) v(dx) \\
C_3(\beta) &= \int_0^1 \left(J(x,m)J(x,n)\right) v(dx)
\end{aligned}
\tag{3.2}
$$

where m, n are counts of aligned DNA sequences from two daughter population and

$$
v(dx) = \theta \frac{s(1) - s(x)}{s(1) - s(0)} m(dx). \tag{3.3}
$$

Suppose that a new nucleotide mutation in a population has population frequency of y. For n chromosomes from that population, then y^n is the probability that the sample is monomorphic of mutant type, $(1-y)^n$ is the probability that the sample is monomorphic of wild type and the probability that the sample is polymorphic is $1 - y^n - (1-y)^n$.

Given a constant γ within each genetic locus, the expected values of the six entries of the DOHRS table are given by:

$$
E(K) = \frac{\theta}{s(1)} \Lambda_1(\gamma, t, m, n) \tag{3.4}
$$

$$
E(O) = \frac{\theta}{s(1)} \Lambda_2(\gamma, t, m, n) \tag{3.5}
$$

$$
E(H) = \frac{\theta}{s(1)} \Lambda_3(\gamma, t, m, n) \tag{3.6}
$$

where

$$\begin{aligned}\Lambda_1(\gamma,t,m,n) = &\int_0^1 \{[I(x,m)K(x,n) + I(x,n)K(x,m) + x^n + x^m] \\ &- \int_0^1 (x^n + x^m)p(t,x,y)m(dy)\}[s(1) - s(x)]m(dx) \\ &+ 2(t - \int_0^t \tilde{P}_0(T_1 \leq u)du)\end{aligned} \tag{3.7}$$

$$\begin{aligned}\Lambda_2(\gamma,t,m,n) = &\int_0^1 \{[J(x,m) + J(x,n) - 2J(x,m)J(x,n) + \\ &2 - x^n - x^m - (1-x)^n - (1-x)^m] \\ &- \int_0^1 (2 - y^n - y^m - (1-y)^n - (1-y)^m)p(t,x,y)m(dy)\} \\ &[s(1) - s(x)]m(dx)\end{aligned} \tag{3.8}$$

$$\Lambda_3(\gamma,t,m,n) = \int_0^1 J(x,m)J(x,n)[s(1) - s(x)]m(dx) \tag{3.9}$$

3.3 Random Effect Bayesian Model

The fixed-effect model assumes that the within-locus selection coefficient γ_i, at locus i, is a constant and follows a normal distribution $N(\mu_r, \sigma_b)$. It is biologically unrealistic to employ an constant selection coefficient to all new mutations that could possibly become fixed or polymorphic at a genetic locus. Here we propose a random effect model by assuming within-locus selection y is also normally distributed with mean γ_i and a global variance σ_w. Without changing the original setting the mean γ_i

varies across loci as a normal $N(\mu_\gamma, \sigma_b)$ in which both parameters are considered to be global parameters. The expected values of the counts due to replacement mutations become conditional expectations, such as $E(K_r|\gamma_i, \sigma_w) = \frac{\theta}{S(1)}\Lambda_1(y|\gamma_i)$, $E(O_r|\gamma_i, \sigma_w) = \frac{\theta}{S(1)}\Lambda_2(y|\gamma_i)$ and $E(H_r|\gamma_i, \sigma_w) = \frac{\theta}{S(1)}\Lambda_3(y|\gamma_i)$. Given that $E(K_r) = E[E(K_r|\gamma_i)]$, we calculate the expected numbers of replacement polymorphisms and fixed differences as

$$\begin{aligned}
E(K_r) &= \int_{-\infty}^{+\infty} E(K_y|\gamma_i)N(\gamma_i, \sigma_w)dy \\
&= \frac{\theta_r}{s(1)}\int_{-\infty}^{+\infty} \Lambda_1(y, t, m, n)N(\gamma_i, \sigma_w)dy \quad = \frac{\theta_r}{s(1)}\Lambda_1^\star(\gamma_i, t, m, n) \\
E(H_r) &= \int_{-\infty}^{+\infty} E(H_y|\gamma_i)N(\gamma_i, \sigma_w)dy \\
&= \frac{\theta_r}{s(1)}\int_{-\infty}^{+\infty} \Lambda_2(y, t, m, n)N(\gamma_i, \sigma_w)dy \quad = \frac{\theta_r}{s(1)}\Lambda_2^\star(\gamma_i, t, m, n) \\
E(O_r) &= \int_{-\infty}^{+\infty} E(O_y|\gamma_i)N(\gamma_i, \sigma_w)dy \\
&= \frac{\theta_r}{s(1)}\int_{-\infty}^{+\infty} \Lambda_3(y, t, m, n)N(\gamma_i, \sigma_w)dy \quad = \frac{\theta_r}{s(1)}\Lambda_3^\star(\gamma_i, t, m, n)
\end{aligned} \tag{3.10}$$

Next, we develop a hierarchical Bayesian framework to the time-dependent random effect PRF model to make statistical inference about divergence and selection between two related species. In contrast to maximum likelihood estimation (MLE), Bayesian method shares information from all loci and hence becomes desirable for multilocus analysis. Although most parameters are governed by information from each individual

locus, species divergence time t_{div},mean selection coefficient μ_γ, between loci variance σ_b and within locus variance σ_w are "global" parameters that make use of information from all loci.

For computational convenience we use conjugate priors for parameter and propose the prior distributions:

$$\begin{aligned}
\theta_s &\sim \Gamma(\alpha_s, \beta_s) \\
\theta_r &\sim \Gamma(\alpha_r, \beta_r) \\
t_{div} &\sim U(0, t_{max}) \\
(\mu_\gamma, \sigma_b) &\sim NG(\mu, \sigma^2 | \mu_0, n_0, \alpha_0, \beta_0) \\
\sigma_w &\sim U(0, \sigma_{Max})
\end{aligned} \tag{3.11}$$

where $NG(\mu, \sigma^2)$ is the inverse-gamma-normal conjugate prior The full likelihood function becomes:

$$\begin{aligned}
&L(\mu_r, \sigma_b^2, \sigma_w^2, t_{div}, K_s, K_r, O_s, O_r, H_s, H_r) \\
&= \prod_{i=1}^{N} \Big\{ \phi(\gamma_i | \mu_\gamma, \sigma_b) \Gamma(\theta_{s,i} | \alpha_s, \beta_s) \Gamma(\theta_{r,i} | \alpha_r, \beta_r) \\
&\qquad \times Poi1(\theta_{s,i}, 0, 0, t_{div}, K_{s,i}, m_i, n_i) Poi2(\theta_{s,i}, 0, 0, t_{div}, O_{s,i}, m_i, n_i) \\
&\qquad \times Poi3(\theta_{s,i}, 0, 0, t_{div}, H_{s,i}, m_i, n_i) Poi1(\theta_{r,i}, \gamma_i, \sigma_w, t_{div}, K_{r,i}, m_i, n_i) \\
&\qquad \times Poi2(\theta_{r,i}, \gamma_i, \sigma_w, t_{div}, O_{r,i}, m_i, n_i) Poi3(\theta_{r,i}, \gamma_i, \sigma_w, t_{div}, H_{r,i}, m_i, n_i) \Big\} \\
&\times \Gamma(\frac{1}{\sigma_b^2} | \alpha_0, \beta_0) \phi(\mu_\gamma | \mu_0, \frac{\sigma_b}{\sqrt{n_0}}) U(t_{div} | 0, t_{max}) U(\sigma_b | 0, \sigma_{wmax})
\end{aligned} \tag{3.12}$$

Due to the complexity of the full likelihood function, it is impossible to handle the question by using one-layer Markov chain Monte Carlo(MCMC) simulation technique. A multi-level MCMC simulation updating strategy was proposed by Bustamante et al. [8]. The trajectories of parameters in each iteration were updated by two types of sampling methods. First is the Metroplis-Hasting method. It compares a likelihood ratio of posterior probabilities that are evaluated by a proposed value of a parameter drawn from a given distribution and the current value, and accept the proposed value if the ratio is greater than a random number from $[0, 1]$[26, 34]. If the full conditional posterior distribution is a known distribution, the Gibbs sampler can directly sample a new value of a parameter from that distribution[21]. It can be shown that the Gibbs sampler is a special case of the Metroplis-Hasting method in which the likelihood ratio is always 100%. Given priors and full likelihood, $\theta_{s,i}$, $\theta_{r,i}$ and hyperparameter (μ_γ, γ_b) are updated by the Gibbs sampling, and σ_w, t_{div}, γ_i are updated by the Metroplis-Hasting method. To implement this model, the "uninformative" prior parameters were applied to impose minimum amount of "prior" knowledge(or artificial assumptions) on the unknown parameters. In our case, we set $\alpha_0 = \beta_0 = \alpha_s = \beta_s = \alpha_r = \beta_r = 0.001, n_0 = 1$, and $t_{max} = 100$, $\sigma_{max} = 10$. Our updating strategies are listed below.

Updating γ_i

In the full likelihood expression (3.12) there are one normal density and three Poisson mass functions involve the parameters γ_i and hence its condition dis-

tribution given other parameters is

$$\begin{aligned}\pi_{\gamma_i}(.) &= C_1\{\phi(\sigma_i|\mu_\gamma,\sigma_b^2)Poi_1(\theta_{r,i},\gamma_i,K_{r,i})Poi_2(\theta_{r,i},\gamma_i,O_{r,i})Poi_3(\theta_{r,i},\gamma_i,H_{r,i}) \\ &= C_2\{\phi(\sigma_i|\mu_\gamma,\sigma_b^2)exp(-\theta_{r,i}\Lambda_i^\star(\gamma,t_{div}))\Lambda_{1i}^\star(\gamma,t_{div})^{K_{r,i}}\Lambda_{2i}^\star(\gamma,t_{div})^{O_{r,i}}\Lambda_{3i}^\star(\gamma,t_{div})^{H_{r,i}}\end{aligned} \tag{3.13}$$

where $\Lambda_i^\star(\gamma,t_{div}) = \Lambda_{1i}^\star(\gamma,t_{div}) + \Lambda_{2i}^\star(\gamma,t_{div}) + \Lambda_{3i}^\star(\gamma,t_{div})$

We update γ_i at each loci by a random-walk Metropolis algorithm. At each time step t we propose a new $\gamma_i^{'}$ by uniformly picking a value in $(\gamma_{i,t} - h_\gamma, \gamma_{i,t} + h_\gamma)$, $h_\gamma = 6$ in our case. We accept this value as a new γ_i value with probability $\min\{1, \frac{\pi_i(\gamma_i')}{\pi_i(\gamma_{i,t})}\}$, or otherwise keep the original value unchanged.

Updating $\theta_{s,i}$ and $\theta_{r,i}$

Th conditional distributions of $\theta_{s,i}$ and $\theta_{r,i}$ given other parameters have closed formulas of gamma distribution and hence Gibbs sampler can be used to draw next step state

$$\begin{aligned}\theta_{s,i} &\approx \Gamma(\alpha_s + K_{s,i} + O_{s,i} + H_{s,i}, \beta_s + \Lambda^\star(0,t_{div})) \\ \theta_{r,i} &\approx \Gamma(\alpha_r + K_{r,i} + O_{r,i} + H_{r,i}, \beta_r + \Lambda^\star(\gamma_i,t_{div})).\end{aligned} \tag{3.14}$$

Updating t_{div}

All Poisson terms and the prior $U(t_{div}|0,t_{max})$ in the full likelihood contain t_{div}

and hence the conditional distribution of t_{div} can be written as

$$\begin{aligned}\pi_t(.) =& C_1 U(t|0, t_{max}) \prod_{i=1}^{N} \{Poi_1(\theta_{s,i}, 0, K_{s,i}) Poi_2(\theta_{s,i}, 0, O_{s,i}) \\ & Poi_3(\theta_{s,i}, 0, H_{s,i}) Poi_1(\theta_{r,i}, \gamma_i, K_{r,i}) \\ & Poi_2(\theta_{r,i}, \gamma_i, O_{r,i}) Poi_3(\theta_{r,i}, \gamma_i, H_{r,i})\}\end{aligned} \tag{3.15}$$

with the ith factor in the product is given by

$$\begin{aligned}& C_{1i} exp[-(\theta_{s,i}\Lambda(0,t) + \theta_{r,i}\Lambda(\gamma_i, t))] \\ & \Lambda_{1i}(\gamma, t_{div})^{K_{r,i}} \Lambda_{2i}(\gamma, t_{div})^{O_{r,i}} \Lambda_{3i}(\gamma, t_{div})^{H_{r,i}} \\ & \Lambda_{1i}(0, t_{div})^{K_{s,i}} \Lambda_{2i}(0, t_{div})^{O_{s,i}} \Lambda_{3i}(0, t_{div})^{H_{s,i}}\end{aligned} \tag{3.16}$$

A newly proposed t'_{div} is drawn uniformly from $(t_{div,t} - h_t, t_{div,t} + h_t)$ with $h_t = 0.7$. Similar to the updating procedure as γ_i we will accept t'_{div} based on the likelihood ratio.

Updating μ_γ and σ_b

We update μ_γ and σ_b from an inverse-gamma-normal distribtuion by first sampling σ_b from

$$\frac{1}{\sigma_b^2} \approx \Gamma(\alpha_a + \frac{N}{2}, \beta_a + \frac{1}{2}\sum_{i=1}^{N}(\gamma_i - \bar{\gamma})^2 + \frac{N n_0 (\frac{1}{N}\sum_{i=1}^{N}\gamma_i - \mu_0)}{2(n_0 + N)}) \tag{3.17}$$

and then sampling μ_γ from

$$\mu_\gamma \approx N(\frac{n_0\mu_0 + \sum_{i=1}^{N}\gamma_i}{n_0 + N}, \frac{\sigma_b^2}{n_0 + N}) \tag{3.18}$$

Updating σ_w

Similar to the situation of γ_i and t_{div} there is no known conditional distribution for σ_w, and it has to be updated by Random-Walk Metropolis algorithm. We set a step $h_s = 0.7$ and sample based on following distribution

$$\pi_{\sigma_w} = \prod_{i=1}^{N} \Big\{ Poi1(\theta_{r,i}, \gamma_i, \sigma_w, t_{div}, K_{r,i}, m_i, n_i) \times Poi2(\theta_{r,i}, \gamma_i, \sigma_w, t_{div}, O_{r,i}, m_i, n_i) \times Poi3(\theta_{r,i}, \gamma_i, \sigma_w, t_{div}, H_{r,i}, m_i, n_i) \times U(\sigma_w|0, \sigma_{max}) \Big\}$$

3.4 *N-ADAM-Mixing* MCMC Sampler

Our initial practice using above method was proven to be unsuccessful in such a way that either the MCMC chain did not converge or converged extremely slow. The reason is that each of the three parameters $(\mu_\gamma, \sigma_b, \sigma_w)$ has a high autocorrelation and it is evidenced by poor mixing behavior of the chain shown in related trace plots and aotucorrelation function(ACF). Such high autocorrelations make proposal values rely heavily on previous values and hence the Markov chain moves slowly through entire parameter space, which causes a slow convergence to true target posterior distribution.

One solution is to run the MCMC simulation as long as we can since the chain will eventually converge as the iteration $t \to \infty$. However, without an explicit estimation this "long enough time" blurs the run time of the MCMC to "infinite", and we have only limited amount of time. A more approachable way is to improve the proposal distribution. Usually a low acceptance rate indicates poor mixing and most draws do not satisfy the target posterior distribution. One cause is that Metropolis sampling steps are too large which decrease probabilities of drawing a acceptable value from the parameter space. Another source is sampling in wrong directions or from a restricted region of entire parameter space. Improving the proposal scale or direction or both can increase the chance that a new draw from proposal distribution will be accepted. Some proposal strategies, such as reparameterization and other adaptive methods[22], have been experimented , but most of which are lack of either rescaling or redirecting proposal distributions.

Haario et al. [24] proposed a adaptive Metropolis(AM) algorithm to tune both step size and spatial orientation of the proposal distribution by assuming a Gaussian proposal distribution. Suppose that at time t a d-dimensional AM Markov chain $X_t \in \mathbf{R}^d$ has been gone through states $X_0, X_1, ..., X_t$. The Gaussian distribution will propose next candidate by setting current Markov chain state X_t as mean and the covariance $s_d C_t$, where variance-covariance matrix C_t is determined by accounting all previous states $X_0, X_1, ..., X_t$ and scaling parameter s_d depends on the dimension d of the vector X_t. Gelman et al. [19] showed that $s_d = (2.38)^2$ is the optimal option for mixing the Metropolis search if both target and proposal distributions are Gaussian, and it should be altered later. Generally it is true but in our practice the target density is so complicated that severely violates the Gaussian assumption. Considered the current state $(\mu_{t-1}, \sigma_{b,t-1}, \sigma_{w,t-1})$ the proposed values are given by

$$\begin{pmatrix} \mu^\star \\ \sigma_b^\star \\ \sigma_w^\star \end{pmatrix} \sim N \left\{ \begin{pmatrix} \mu_{t-1} \\ \sigma_{b,t-1} \\ \sigma_{w,t-1} \end{pmatrix}, C_t \right\}$$

where C_t is

$$C_t = \begin{cases} C_0 & t \leq t_0 \\ S_d Cov(X_0, ..., X_{t-1}) + S_d \varepsilon I_d & t \geq t_0 \end{cases}$$

The acceptance probability given by the Metropolis algorithm is

$$\alpha\{X_t, X^\star\} = \min\{1, \frac{\pi(X^\star)}{\pi(X_t)}\} \tag{3.19}$$

where X_t is current state and $X^\star$ is the proposed value.

The empirical variance-covariance matrix could be updated with a recursive algorithm to reduce computational cost, such as

$$C_t = Cov(X_1, ..., X_t) = \frac{t-2}{t-1} C_{t-1} + \frac{1}{t-1}\left\{X_t X_t^T + (N-1)\bar{X}_{t-1}\bar{X}_{t-1}^T - \bar{X}_t \bar{X}_t^T\right\}$$

$$\bar{X}_{t-1} = \frac{1}{t-1}\sum_{i=1}^{t-1} X_i \qquad \bar{X}_t = \left\{(t-1)\bar{X}_{t-1} + X_t\right\}/t$$

In an application of this method to our data set, we combine $(\mu_\gamma, \sigma_b, \sigma_w)$ as a $3D$ joint distribution and update them using an empirical covariance matrix. Accordingly we have to change the $3D$ density as

$$\begin{aligned}\pi_{joint}(.) = \prod_{i=1}^{N} \Bigg\{ & \phi(\gamma_i|\mu_\gamma, \sigma_b) \times Poi1(\theta_{r,i}, \gamma_i, \sigma_w, t_{div}, K_{r,i}, m_i, n_i) \\ & \times Poi2(\theta_{r,i}, \gamma_i, \sigma_w, t_{div}, O_{r,i}, m_i, n_i) Poi3(\theta_{r,i}, \gamma_i, \sigma_w, t_{div}, H_{r,i}, m_i, n_i) \Bigg\} \\ & \times U(\sigma_b|0, \sigma_{max}) U(\mu_\gamma|\mu_{min}, \mu_{max}) U(t_{div}|0, t_{max}) U(\sigma_w|0, \sigma_{max})\end{aligned} \tag{3.20}$$

For the first $50,000$ iterations, we fixed the variance-covariance matrix at $C_0 = \begin{pmatrix} 50 & 0 & 0 \\ 0 & 25 & 0 \\ 0 & 0 & 25 \end{pmatrix}$. In our trails this method did reduce autocorrelations but the sampling

efficiency is still low due to the high correlation among $(\mu_\gamma, \sigma_b, \sigma_w)$.

A high dimension MCMC, like our case, introduces certain significant correlations among parameters. The searching path is then dominated by some of the parameters due to the high correlation and will be limited to a small region of the parameter space. For example, there are two random variables X_1, X_2 follow a joint normal distribution with a variance-covariance matrix $C = \left(\begin{smallmatrix} 1.5 & 1.4 \\ 1.4 & 1.5 \end{smallmatrix}\right)$. Obviously they are highly correlated. If we use C as our covariance of the Gaussian proposal distribution, the sampling region will be a flat-needle-shape ellipse. The Metropolis sampler will only go through the region around a 45-degree axes within an extremely narrow width, as shown in the top panel of Figure 3.1.

Similarly, the high correlation among $(\mu_\gamma, \sigma_b, \sigma_w)$ compressed the searching path to a narrow needle shape area, where the pinpoints region are remarkably smaller than the middle region of the needle. Even after the AM tuning these pinpoints can hardly be touched, and the proposal distribution will be trapped at the middle region. Such behavior will cause a low acceptance rate due to the fact that inappropriate searching path does not explore true area of sample space and sample draws wander around the middle region.

Bai [4] proposed an adaptive directional Metropolis-within-Gibbs (ADMG) algorithm that can adjust both random sample direction and scale "componentwisely" with a Metropolis-within-Gibbs sampler. A singular value decomposition(SVD) is performed on the empirical covariance matrix, and then orthonormal vectors from SVD are used as sampling directions. Referring to "componentwisely", this algorithm

updates each parameter following one of the sampling directions from SVD and with a jumping scale that is tuned based on average acceptance rates from previous 100 steps.

In our practice we adapted both AM and ADMG algorithms to develop an *Adaptive directional Adaptive Metropolis* ($ADAM$) algorithm to ensure the efficiency and convergence of our MCMC. Rather than componentwise we still update $(\mu_\gamma, \sigma_b, \sigma_w)$ as a $3D$ vector.

For the previous $C = \left(\begin{smallmatrix} 1.5 & 1.4 \\ 1.4 & 1.5 \end{smallmatrix}\right)$, we conduct a singular value decomposition (SVD) on this matrix which projects the two variables into an orthogonal space by

$$C = D\Sigma U \tag{3.21}$$

where Σ is a 2×2 diagonal matrix and D and U are unitary with $U = D^T$. The columns of U are a set of orthonormal vectors that span a vector space. Usually D and U are regarded as rotation matrices which rotate coordinates into an orthogonal space and Σ as a scaling matrix which describes the lengths of each axes after rotation.

Based on the SVD,

$$D = \begin{pmatrix} \text{-0.7071068} & \text{-0.7071068} \\ \text{-0.7071068} & 0.7071068 \end{pmatrix}$$

$$\Sigma = \begin{pmatrix} 2.9 & 0 \\ 0 & 0.1 \end{pmatrix} \tag{3.22}$$

$$U = D^T$$

Performing this orthogonal transformation on X_1 and X_2 gives us a new set of random variables Y_1 and Y_2 as follows

$$\begin{pmatrix} Y_1 \\ Y_2 \end{pmatrix} = U \begin{pmatrix} X_1 \\ X_2 \end{pmatrix} \tag{3.23}$$

After the transformation, the proposal distribution will sample next candidates Y_1' and Y_2' based on Y_1 and Y_2 instead of X_1 and X_2 along with variance-covariance matrix Σ.

$$\begin{pmatrix} Y_1' \\ Y_2' \end{pmatrix} \sim N\left\{ \begin{pmatrix} Y_1 \\ Y_2 \end{pmatrix}, \Sigma \right\} \tag{3.24}$$

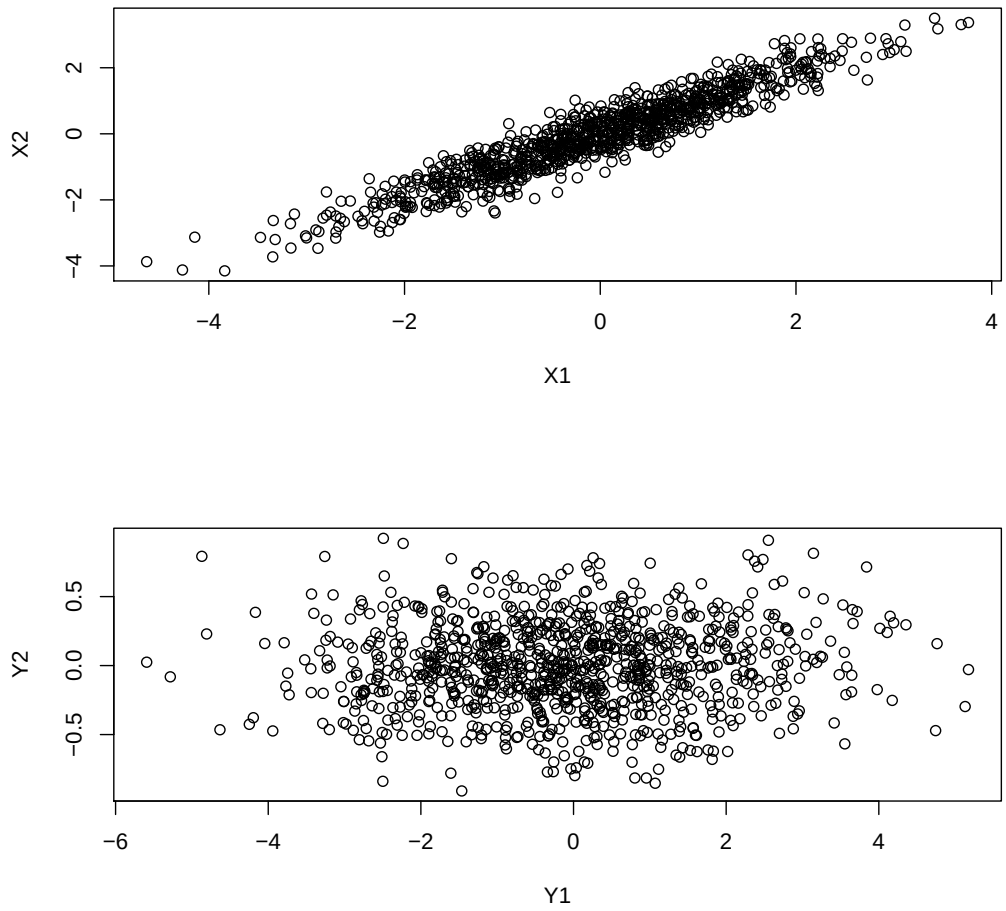

Figure 3.1: Sampling Region

This method projects the original sampling region to a significantly expanded space, and provides better chance to draw more "satisfied" candidates(bottom panel of Figure 3.1). Thus it improves the efficiency of the proposal distribution and moves the MCMC towards the target density more quickly.

While updating $(\mu_\gamma, \sigma_b, \sigma_w)$ we apply a $\delta(n)\Sigma$ as covariance matrix, where $\delta(n) =$

$exp(2 * d * (\delta(n)^{(k)} - 0.3))$ is a jumping scale. Here, $\delta(n)^{(k)}$ is interpreted as 100–step average acceptance rate at "k_{th}" batch of iterations. We don't have to update the 100–step average acceptance rate $\delta(n)^{(k)}$ in every iteration. Alternatively it will be updated for every 100 iterations, that is after "k^{th}" 100 iterations it will be reset as $\delta(n)^{(k)} = \frac{1}{100}\Sigma_{n=k*100+1}^{k*100+100}\alpha_n$, where $\alpha_n = \min\{1, \frac{\pi(Y^\star)}{\pi(X_t)}\}$ is the acceptance rate for each iteration. We describe the $ADAM$ algorithm on $X = (\mu_\gamma, \sigma_b, \sigma_w)$ as follow:

- Step 1

  Perform singular value decomposition on $\Sigma_t = D\Sigma U$. Set $Y = U^T X_t$ where $X_t = (\mu_t, \sigma_{b,t}, \sigma_{w,t})$

- Step 2

  Sample $Y^\star$ from $N\{Y, \delta(n)\Sigma\}$

- Step 3

  Revise the transformation $X^\star = \left(U^T\right)^{-1} Y^\star$

- Step 4

  Calculate the acceptance rate $\alpha = \min\{1, \frac{\pi_{joint}(X^\star)}{\pi_{joint}(X_t)}\}$. Set $X_{t+1} = X^\star$ with probability α, otherwise $X_{t+1} = X_t$

However fixing the initial matrix at $C_0 = \begin{pmatrix} 50 & 0 & 0 \\ 0 & 25 & 0 \\ 0 & 0 & 25 \end{pmatrix}$ still concerns us since it is purely our guess from biological senses and all other references tell us no information of choosing such constant matrix. Even from Bayesian perspective this process is not safe. With all knowledge of former runs we develop a process that quantitatively exhibits behaviors of each datasets. A natural implementation is to use original method

to start the MCMC run, and then calculate variance-covariance matrix based on that. We name the new method as *N-ADAM-Mixture* algorithm.

- Initial

  Run the MCMC using the original method(without any adaptions) for $50,000$ and calculate the empirical covariance matrix C_0.

- Fixed

  Run the *ADAM* method for another $100,000$ iterations using the fixed covariance matrix C_0.

- Full Adaption

  Fully adapt *ADAM*, update empirical covariance matrix C_t recursively for every iteration.

3.4.1 Ergodicity

The *N-ADAM-Mixture* chain is essentially an *ADAM* chains. Because we proposed a different approach to "guess" what is the best C_0, but did not change the updating strategy after that. However a *ADAM* chain is no longer *Markovian*. This is because the empirical variance-covariance matrix uses the cumulative information on all previous states, hence the transition kernel depends on $X_0, X_1, ..., X_n$ as $P(X_{n+1}|X_0, X_1, ..., X_n)$ instead of just X_n .

The orthogonal transformation from Y to X is solely to expand our sampling

region. In fact $X_{n+1} = U^{-1}Y \sim N(X_n, \Sigma_n)$, the *empirical effects* Σ_n can still be viewed as from X_n to X_{n+1} directly.

For a Markov chain $\{X_t\}$ we define $\pi(.)$ as its target probability distribution on a state space $\mathcal{X}$. Let $\{P_\gamma\}$ be a collection of Markov chain kernels that have the same stationary distribution $\pi(.)$. Updating from X_n to X_{n+1} we have

$$P(X_{n+1}|X_n, \Gamma_n, X_{n-1}, \Gamma_{n-1}, ..., X_0, \Gamma_0) = P_{\Gamma_{n+1}}(x, .) \text{ for } n \geq 0 \tag{3.25}$$

where $\Gamma_n \in \mathcal{Y}$ is a random index chosen at the n^{th} step. Roberts and Rosenthal [40] proved that

$$\begin{aligned} &\lim_{n\to\infty} \sup_{A\subset\mathcal{X}} ||P(X_n \in A) - \pi(A)|| = 0 \\ &\lim_{n\to\infty} \frac{1}{n}\Sigma_{i=1}^{n} g(X_i) = \pi(g) \text{ for all bounded } g : \mathcal{X} \to \mathbf{R} \end{aligned} \tag{3.26}$$

given that both *Diminishing*(or *Vanishing*) *Adaption* condition

$$\lim_{n\to\infty} \sup_{x\subset\mathcal{X}} ||P_{\Gamma_{n+1}}(X, .) - P_{\Gamma_n}(X, .)|| = 0 \text{ in probability} \tag{3.27}$$

and *Containment*(*Bounded Convergence*) condition

$$\{M_\epsilon(X_n, \Gamma_n) = \inf\{n \geq: ||P_{\Gamma_n}(X_n, .) - \pi(x, .)|| \leq \epsilon\}\}_{n=0}^{\infty} \text{ is bounded in probability, } \epsilon > 0 \tag{3.28}$$

hold. This theorem provides a powerful "detector" to verify whether or not X_n preserves its ergodicity and hence $P_{\Gamma_n \in \mathcal{Y}}(x, .)$ will converge to the stationary distribution

$\pi(.)$.

The *ADAM* adaption uses $X_0, X_1, ..., X_{n-1}$ to estimate an empirical covariance matrix as the variance-covariance matrix for the Gaussian proposal distribution at n^{th} step. The amount of change made in the estimated variance-covariance matrix at step n is only order of $O(\frac{1}{n})$. Our scale parameter $\delta(n)$ can be expressed as $\delta(n) = C + \theta(n)$ where C is a constant and $\theta(n)$ is an adjusting parameter at step n. As $n \to \infty$, both $O(\frac{1}{n})$ [40]and $\theta(n)$ are going to be 0[4]. It satisfies diminishing condition.

In particular, to avoid wasting time for parameters wandering at some states that are non-biological meaning, we defined the upper and lower boundaries for $\mu_\gamma \in [-20, 20]$, both σ_b and $\sigma_w \in [0, 10]$. Thus Haario et al. [24] claimed that empirical covariance matrix Σ_n is bounded such that $\Gamma_n \in \mathcal{Y}$ is compact. Therefore $\mathcal{X} \times \mathcal{Y}$ is compact and the containment condition holds[5, 39]. Hence the ergodicity holds for *N-ADAM* algorithm.

3.5 Numerical Approximation and Parallel Computing

In our diffusion approximation, the integration in the following form

$$\int_0^1 f(y)p(x,y,t)m(dy) \tag{3.29}$$

is a solution of a parabolic partial differential equation(PDE) with certain boundary conditions and an initial condition

$$\begin{aligned} &\mu_{tt}(x,t) = a(x)\mu_x(x,t) + b(x)\mu_{xx}(x,t) \\ &\mu(0,t) = \mu(1,t) = 0, \qquad\qquad \mu(x,0) = f(x) \quad \text{for} \quad 0 \le x \le 1 \end{aligned} \tag{3.30}$$

A classic *Crank–Nicolson(CN)* method can implicitly numerically solves this PDE with a fixed time step Δt on $[0,t]$ and a fixed space step Δx on $[0,1]$. Another type of integration which occurs multiple times in our model is $\int_0^1 g(x)m(dx)$ and it can be evaluated numerically by Gaussian-Legendre approximation with $m(dx) = \frac{e^{\gamma x}}{x(1-x)}dx$ for $\gamma \ge 0$. We can substitute $g(x)\frac{e^{\gamma x}}{x(1-x)}$ with $f(x)$, and derive the solution as

$$\int_0^1 f(x)x(1-x)dx = \frac{1}{2}\Sigma_{k=1}^{n} w(\xi_k) f(\frac{1}{2}\xi_k + \frac{1}{2}) \tag{3.31}$$

where abscissas ξ_k for $k = 1, 2, ..., n$ are roots of the Legendre polynomials $P_n(x)$ and corresponding weight functions $w(\xi_k)$ are obtained by solving a system of equations. To reduce the computational cost without sacrificing accuracy we set $n = 10$.

The most time consuming part in our calculation is to numerically evaluate a

two-layer integration such as $\int_0^1 \mu(x,t)m(dx)$ which shows up frequently in those Λ functions. Notice that the $\mu(x,t)$ satisfies the PDE (3.30). It then requires a Crank-Nicolson method to corporate with the Gaussian-Legendre integration so that space step x_i of CN method is no longer fixed. That is, the grid mesh of CN method on x space is not uniform any more. We have to numerically solve the PDE on $x_i = \xi_k$ for $k = \{1, 2, ..., 10\}$ according to the Gaussian-Legendre method. Hence a Nonuniform grid CrankNicolson method is applied to these functions. Under this circumstance we evaluate a two-layer function as

$$\int_0^1 \mu(x,t)m(dx) = \frac{1}{2}\Sigma_{k=1}^{10} w(\xi_k)\mu(\frac{1}{2}\xi_k + \frac{1}{2}, t)\frac{e^{\gamma\xi_k}}{\xi_k(1-\xi_k)} \tag{3.32}$$

for a time $t > 0$. That is, at a fixed time t we numerically solve a parabolic PDE for 10 times in order to retrieve an answer of this two-layer function.

Both Nonuniform grid CrankNicolson method and Gaussian-Legendre method are implemented with R API, and later can be called in R.

For our random effect model, there is another type of integration needs to be evaluated numerically. That is

$$E(K_r) = E[E(K_r|\gamma_i)] = \frac{\theta}{S(1)}\int_{-\infty}^{+\infty} \Lambda_1(y|\gamma_i)N(y|\gamma_i, \sigma_w)dy \tag{3.33}$$

where Λ_1 is the Λ function derived from fixed effect model. *Gauss-Hermite* method is designed to deal with integration over interval of $(-\infty, +\infty)$. Let $E(K_r) =$

$\frac{\theta_r}{s(1)}\Lambda_1^\star(\gamma_i, \sigma_w, t)$ and the numerical process can be described as follow

$$E(K_r) = \frac{\theta}{S(1)} \int_{-\infty}^{+\infty} \Lambda_1(y|\gamma_i) \frac{1}{\sqrt{2\pi}\sigma_w} e^{-\frac{(y-\gamma_i)^2}{2\sigma_w^2}} dy$$

Set $x = \frac{y-\gamma_i}{\sigma_w} \sim \phi(0,1)$

since $\Lambda_1(\gamma_i + x \cdot \sigma_w|\gamma_i) = \Lambda_1(\gamma_i + x \cdot \sigma_w)$

$$\Rightarrow \frac{\theta}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} \Lambda_1(\gamma_i + x \cdot \sigma_w) e^{-\frac{x_i^2}{2}} dx$$

Set $\frac{x}{\sqrt{2}} = y$

$$\Rightarrow \frac{\theta}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} \Lambda_1(\gamma_i + \sqrt{2} \cdot y \cdot \sigma_w) e^{y^2} dy$$

$$\Rightarrow \Lambda_1^\star(\gamma_i, \sigma_w, t) = \frac{1}{\sqrt{\pi}} \int_{-\infty}^{+\infty} \Lambda_1(\gamma_i + \sqrt{2} \cdot y_j \cdot \sigma_w) e^{y_j^2} dy_j$$

$$= \frac{1}{\sqrt{\pi}} \sum_{j=1}^{10} \Lambda_1(\gamma_i + \sqrt{2} \cdot y_j \cdot \sigma_w) w(y_j)$$

where y_j and $w(y_j)$ are abscissas and weights respectively. Similar to Gauss-Legendre method we select 10 pairs of abscissas and weights. Same method applied to approximating the values of $E(H_r)$ and $E(O_r)$.

Multiple occurrence of these two-layer integrations and expected values from random effect model require us to solve the PDE for over 2000 times just in one single MCMC iteration. This computation intensity prevents us from implementing the Bayesian framework developed in the random effect model by using MCMC simulation. Because such MCMC simulation needs at least one million iterations before it can be considered to converge and it takes over 300 days based on rough estimation.

To increase the computing speed, we have to redesign the program to suit a parallel

computing architecture. At a given genetic locus i, estimates of functions $\Lambda^{\star}_{1,i}, \Lambda^{\star}_{2,i}, \Lambda^{\star}_{3,i}$ are independent of each other so that each function consumes an independent I/O. All genes in our data set are also independent of each other which ensures an independent I/O for each gene. Hence, within a single MCMC iteration function values of $\Lambda^{\star}_{1,i}, \Lambda^{\star}_{2,i}, \Lambda^{\star}_{3,i}$ at genetic locus i can be independently estimated without affecting the same process at other locus j for $i \neq j$. Thus an parallel computing technique can be implemented in our model and boosts loop speed.

Message Passing Interface (MPI) is a communication protocol to standardize the message-passing system on parallel computers. The MPI defines syntax and semantics of a core set of library routines[12]. Such library routines provide varieties for users to implement message passing programs on parallel computers in Fortran and C, and later script language such as Python, Jave etc. Implementations of MPI mostly use a manager-worker architecture in which multiple processes(workers) are controlled by one process(manager). Therefore, memory management and "message"(data and functions) passing mechanism between a manager and workers are extremely useful for most statistical users. An alternative solution which have been adapted within statistical computation is parallel computing packages under R environment.

R[46] is an open-source programming language and software environment for statistical analysis. It provides a script-like language environment that reduce the complexity for any individuals to perform statistical analysis. Through an extensive amount of packages R can be easily extended to many complicated statistical process. In R, packages are libraries of functions and R user can develop and distribute R

packages at the Comprehensive R Archive Network(CRAN) under suitable licenses. It provides options for almost every aspect of statistical studies and superb graphic ability. However, parallel or high performance computing (HPC) is not natively supported by R environment. Several packages have been developed to compensate this disadvantage. An early effort is Rmpi [48] which provides a low level programming interface on MPI. Without knowing details of MPI implementations R users can access low level MPI functions throught Rmpi functions. But it is only a wrapper of MPI which is still too complicated to be widely used by R users. Especially, users have to have a deep knowledge of mechanism of a manager-worker architecture to span and manage worker processes. *Simple Network of Workstations(snow)* provides wrappers of Rmpi functions. R instances on workers processes are launched through a script ($c < makeCluster()$). It also supports alternative version of high-level apply() function family. *snowfall* package is on top of *snow* package, which provides the top-level wrappers of *snow* functions. *Snowfall* grands user an simple interface to launch a cluster computing by *sfInit()* function without handling R cluster objects. For example, in a cluster provided by the Center for Applied Mathematics and Statistics(CAMS) at UNLV, we can launch a R parallel computing interactive session by *sfInit(parallel=T,cpus=10,type="MPI")*. This function will form a computing group of 11 CPUs including 1 manager and 10 workers under MPI communication protocol. Any data that is acquired by processes on workers can be explicitly passed from manager via *sfExport()* function. In addition, functions that will be executed on workers are loaded on worker nodes by *sfSource()* function, which provides a convenient way

to handle the worker computing by declaring a separated file including all worker functions. Along with high-level *apply()* function family *snowfall* can automatically distribute and execute calculation on different nodes over the cluster. For a cluster that has machines with different calculation speed *sfClusterApplyLB()* can balance this infrastructure by immediately starting a new segment on a node upon completion of it's previously assigned segment.

The OpenMPI is the most widely used MPI implementation these days. However, it is not suitable to R MPI packages. Instead, we used LAMMPI implementation in our program.

To apply a parallel computing scheme to our dates set, we construct a 91×3 table where 91 rows represents 91 genetic loci from two species and 3 columns contains three $\Lambda^\star$ functions. In CAMS cluster system, we distribute 91 rows into 32 CPUs on different four nodes through MPI interface under R environment. The total 32 CPUs consist of one manager and 31 workers who will carry out most of the calculation. Because all three $\Lambda^\star$ functions rely only on three parameters: t_{div}, σ_w and γ_i at the i^{th} locus, updates have to be made upon changes of these three parameters. To minimums communication time between the manager and workers, we proposed the parallel updating scheme as follow:

- Step 0 initialize the 91×3 $\Lambda^\star$ table using the initial values proposed by random distributions(once at the beginning of a run).

- Step 1 propose new $t^\star_{div}$, $\gamma^\star_i$, $(\mu^\star_\gamma, \sigma^\star_b, \sigma^\star_w)$ for $i = 1, 2, ..., 91$

- Step 2 Calculate a new 91×3 $\Lambda^\star$ table based on the proposed $\gamma^\star$, replace i^{th} row of $\Lambda^\star$ table if the i^{th} proposed $\gamma^\star$ has been accepted.

- Step 3 Given the $t^\star_{div}$ calculate a new 91×3 $\Lambda^\star$ table and replace the old $\Lambda^\star$ table with a new $\Lambda^\star$ table if $t^\star_{div}$ has been accepted.

- Step 4 Calculate a new 91×3 $\Lambda^\star$ table based on $(\mu^\star_\gamma, \sigma^\star_b, \sigma^\star_w)$, replace the old $\Lambda^\star$ table with a new $\Lambda^\star$ table if $(\mu^\star_\gamma, \sigma^\star_b, \sigma^\star_w)$ has been accepted.

Step1$\sim$4 are executed within the main loop body, while Step 0 is an initialization procedure that provides some randomly chosen values to start the MCMC process. The MCMC running information is printed to a .Rout file and can be monitored real time. For every $50,000$ iterations all estimated parameters are saved to .rda file for future analysis.

CHAPTER 4

STATISTICAL INFERENCES OF DIVERGENCE AND SELECTION

4.1 Simulation Study

To test the ability of estimating divergence time accurately, we applied our model to two simulated date sets with two extreme time. They are 4.38 for data set 1 and 0.56 for data set 2, where the divergence time are scaled in terms of diffusion time scale. Each of the two sets are simulated to contain 30 genes with parameters $(\mu_\gamma, \sigma_b, \sigma_w)$ taking $(-6.82, 3.78, 2.56)$ for data set 1, and $(9.15, 3.15, 2.37)$ for data set 2 respectively. After the first $250,000$ iterations as a burn-in period, $5,000$ samples are taken every 400 steps to form ten sub chains. The convergence is confirmed by trace plots (Figure 4.1 and 4.2) and Gelman–rubin diagnostic (< 1.1)[20].

For both data sets, the divergence time quickly converged to its true value with slight variation. However, all simulated results tend to overestimate global parameters μ_γ, σ_b and σ_w. For example, the magnitude of the mean selection coefficient $\hat{\mu}_\gamma$ is larger than its corresponding true value, but maintains the same selective direction in the sense that the sign of $\hat{\mu}_\gamma$ stays the same as the given value. Also median estimates of $\hat{\sigma}_w$ in our study are not close to their true values but 95% confidence intervals still cover most true values. The reason is that σ_w is an artificial parameter we implanted into the model to be biologically realistic, but it is lack of data support.

It may require a longer MCMC simulation for the model to capture the true values of σ_w or add more loci into the data set to supply more information about within-locus variation.

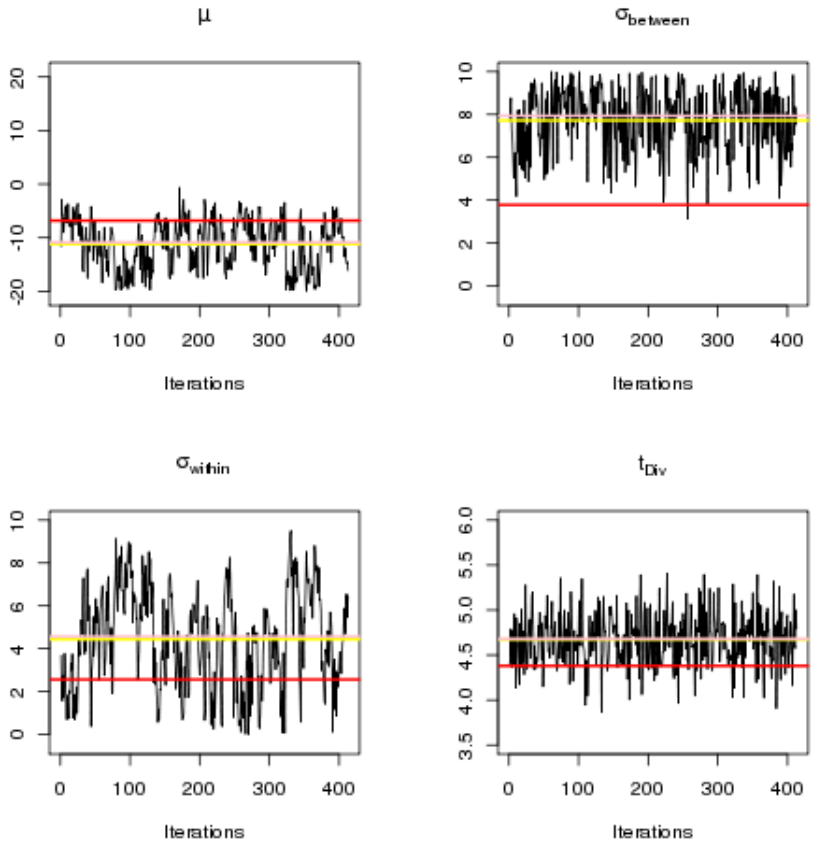

Figure 4.1: Trace plots of data set 1

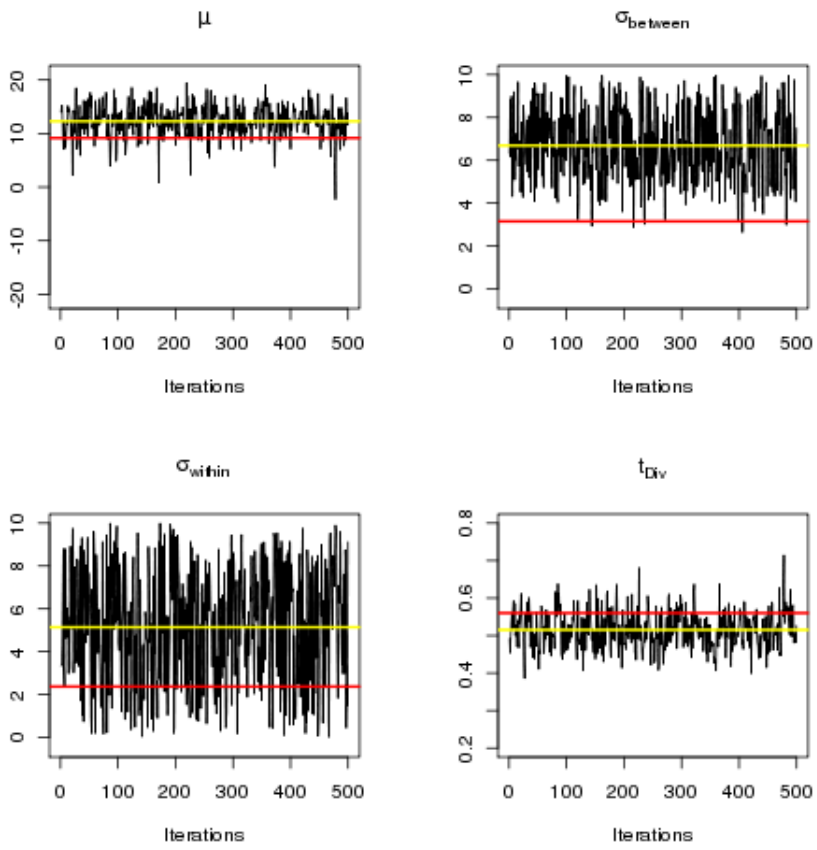

Figure 4.2: Trace plots of data set 2

4.2 Real Data Set

The time-dependent random effect PRF model was applied to the data of Pröschel et al. [37]. The data set consists of 91 autosomal genes in *Drosophila melanogaster* which are collected from Lake Kariba, Zimbabwe. The number of alleles ranges from 7 to 12 are all from coding regions[23]. As a comparison of divergence a single highly inbred line of *Drosophila simulans* is sampled from Chapel Hill, North Carolina[33]. Each of the 91 genes forms a DOHRS table consisting of K_s, O_s, H_s, K_r, O_r, H_r and

total numbers of alleles from each species (n ranges from 7 to 12 and $m = 1$). Thus the data can be viewed as a 91×8 matrix.

After the first $150,000$ burn-in iterations, parameters were estimated from 10 MCMC sub chains that each had 500 samples drawn by every 400 iterations. Both ACF and Gelman–rubin diagnostic (< 1.1) grant the convergence of the MCMC sub chains[20]. We present our results for the four parameters μ_γ, σ_b, σ_w and t_{div} in terms of their means, standard errors,medians, 95% credible intervals and 6-lag autocorrelations and list them all in Table 4.1 and Table 4.2.

		Mean	S.D	Median	G.R.	HPDI(95%)		ACF				
						Lower	Upper	1	2	3	4	5
SubChain 1	μ_γ	-3.68	3.65	-3.81		-10.25	2.42	0.81	0.75	0.68	0.63	0.57
	$\sigma_{between}$	6.28	1.64	6.22		3.44	9.48	0.51	0.47	0.38	0.38	0.35
	σ_{within}	6.19	2.53	6.6		0.8	9.79	0.89	0.82	0.74	0.69	0.65
	t_{Div}	2.7	0.11	2.69		2.48	2.9	0.08	0.04	0.03	0.06	0.03
SubChain 2	μ_γ	-3.18	3.84	-2.71	1.01	-10.7	2.66	0.86	0.79	0.72	0.66	0.61
	$\sigma_{between}$	6.07	1.63	6.06	1.01	3.15	9.35	0.63	0.54	0.48	0.44	0.42
	σ_{within}	5.79	2.69	6.08	1.01	0.44	9.84	0.91	0.84	0.77	0.71	0.65
	t_{Div}	2.69	0.11	2.69	1	2.5	2.91	-0.02	0.01	-0.09	0.06	-0.06
SubChain 3	μ_γ	-3.3	3.77	-2.93	1	-10.27	2.51	0.87	0.79	0.74	0.69	0.63
	$\sigma_{between}$	6.24	1.66	6.2	1	3.35	9.5	0.59	0.5	0.45	0.42	0.37
	σ_{within}	6	2.58	6.25	1	0.93	9.84	0.91	0.84	0.78	0.71	0.65
	t_{Div}	2.69	0.11	2.69	1	2.49	2.91	0.02	-0.01	0.01	-0.06	-0.04
SubChain 4	μ_γ	-3.02	4.17	-2.32	1	-10.27	2.7	0.89	0.83	0.76	0.71	0.67
	$\sigma_{between}$	5.96	1.73	5.93	1	3.07	9	0.68	0.59	0.53	0.45	0.46
	σ_{within}	5.56	2.97	5.71	1.01	0.37	9.72	0.92	0.86	0.8	0.73	0.69
	t_{Div}	2.69	0.11	2.68	1	2.49	2.93	-0.1	0.05	-0.04	-0.01	0.03
SubChain 5	μ_γ	-3.63	3.91	-3.6	1	-10.82	2.52	0.87	0.79	0.74	0.7	0.64
	$\sigma_{between}$	6.27	1.64	6.28	1	3.49	9.41	0.63	0.55	0.51	0.46	0.44
	σ_{within}	6.1	2.71	6.57	1.01	0.57	9.84	0.92	0.85	0.79	0.75	0.69
	t_{Div}	2.68	0.11	2.69	1	2.46	2.91	0.07	0.06	-0.08	0.01	-0.06

Table 4.1: Estimates of μ_γ, σ_b, σ_w and t_{div}

		Mean	S.D	Median	G.R.	HPDI(95%)		ACF				
						Lower	Upper	1	2	3	4	5
SubChain 6	μ_γ	-3.31	3.65	-3.15	1	-10.14	2.71	0.85	0.77	0.71	0.64	0.56
	$\sigma_{between}$	6.22	1.6	6.25	1	3.38	9.25	0.59	0.53	0.46	0.43	0.39
	σ_{within}	6.04	2.51	6.23	1	0.69	9.7	0.9	0.82	0.74	0.67	0.6
	t_{Div}	2.69	0.11	2.68	1	2.47	2.91	-0.02	0.02	0.02	0.01	-0.01
SubChain 7	μ_γ	-3.51	3.9	-3.76	1	-10.11	2.68	0.87	0.8	0.77	0.71	0.68
	$\sigma_{between}$	6.27	1.75	6.29	1	3.3	9.45	0.7	0.62	0.61	0.58	0.55
	σ_{within}	6.04	2.81	6.77	1	0.59	9.72	0.92	0.87	0.83	0.79	0.75
	t_{Div}	2.69	0.11	2.69	1	2.49	2.89	-0.04	0.02	0.05	-0.04	-0.03
SubChain 8	μ_γ	-4.51	3.77	-4.89	1.01	-11.04	1.79	0.85	0.76	0.69	0.65	0.59
	$\sigma_{between}$	6.41	1.54	6.38	1	3.6	9.42	0.59	0.43	0.31	0.29	0.27
	σ_{within}	6.7	2.34	7.17	1.01	1.79	9.87	0.89	0.81	0.74	0.68	0.63
	t_{Div}	2.7	0.11	2.71	1	2.49	2.93	0.02	-0.02	-0.05	0.05	-0.02
SubChain 9	μ_γ	-2.12	3.47	-1.7	1.01	-9.43	2.92	0.85	0.8	0.74	0.68	0.64
	$\sigma_{between}$	5.86	1.6	5.72	1.01	3.35	9.24	0.57	0.51	0.53	0.42	0.4
	σ_{within}	5.17	2.62	5.28	1.01	0.3	9.59	0.89	0.82	0.77	0.73	0.7
	t_{Div}	2.68	0.11	2.67	1	2.46	2.91	0.04	0.01	0.05	0.05	0.01
SubChain 10	μ_γ	-2.87	3.95	-2.81	1.01	-9.71	2.68	0.9	0.85	0.8	0.75	0.7
	$\sigma_{between}$	5.94	1.63	6	1.01	3.27	9.09	0.66	0.56	0.52	0.47	0.47
	σ_{within}	5.57	2.87	6.16	1.01	0.39	9.76	0.92	0.86	0.81	0.75	0.71
	t_{Div}	2.67	0.11	2.67	1	2.48	2.89	0.05	0.03	-0.03	0.01	-0.04

Table 4.2: Estimates of μ_γ, σ_b, σ_w and t_{div}(continued)

All estimates discussed below are based on the values generated from the last sub chain. In the diffusion time scale the point estimates (median) and 95% creditable intervals of the global parameters are mean of selection coefficients $\mu_\gamma = -2.81$ with $(-9.71, 2.68)$, between-locus standard deviation $\sigma_b = 6.00$ with $(3.27, 9.09)$, within-locus standard deviation $\sigma_w = 6.16$ with $(0.39, 9.76)$ and species divergence time $t_{div} = 2.67$ with $(2.48, 2.89)$. The mean selection coefficients μ_γ implies that the selection coefficient γ at each locus tends towards negative value due to model assumption that it is normally distributed with mean μ_γ. Our results support the viewpoint that most replacement mutations are deleterious[8, 16, 17, 41, 43, 44]. However, comparing with $\mu_\gamma = -5.7$ estimated in Sawyer et al. [44], our estimate is only half of that value and more lean towards so-call "mild deleterious". The estimated median selection coefficients and their 95% creditable intervals for the 91 genes are plotted in Figure 4.3 and sorted in an ascending order. In the Figure 4.3, only 24 out of the 91 genes have positive selection coefficients γ. Overall, the magnitude of the selection coefficients is small, for example, 36% of the replacement mutations have $\gamma > -1$, 76% have $\gamma > -5$, and 93% have $\gamma > -9$.

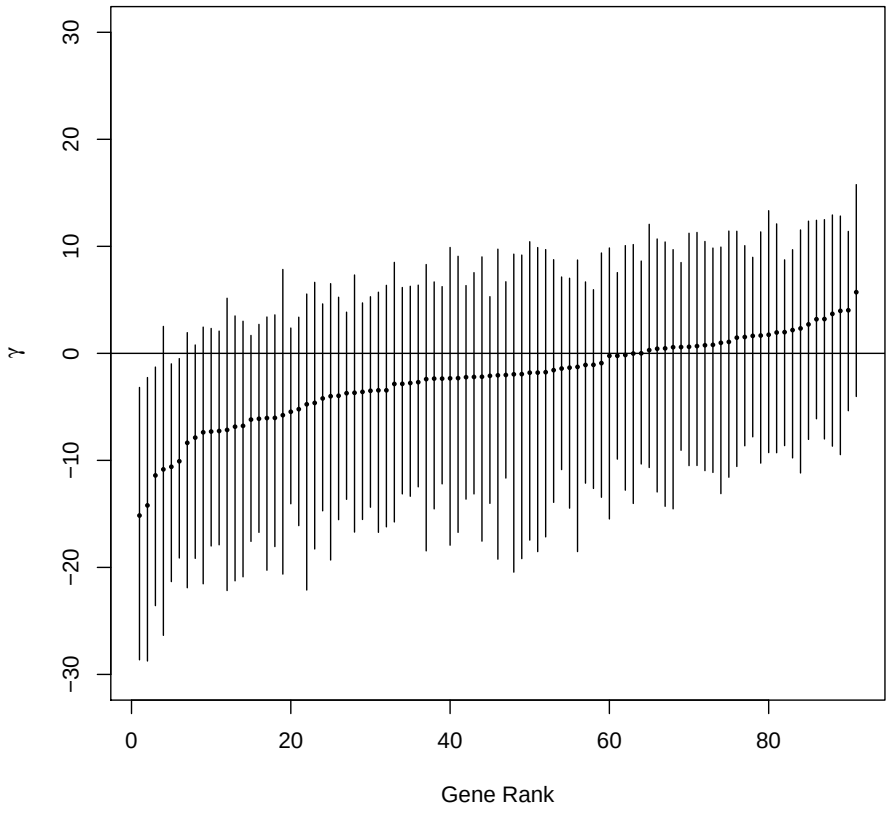

Figure 4.3: Sorted selection coefficient γ for the 91 genes

Adapting the haploid effective population size of $N_e = 0.645$ million, the estimate of divergence time $t_{div} = 2.67 \pm 0.11$ implies that the divergence between *D.melanogaster* and *D.simulans* happened 1.72 million year ago. It is consistent with Amei's estimate of $t_{div} = 2.61$ under the time-dependent fixed effect model[3], which suggests that in terms of divergence time estimation, the time-dependent fixed effect model is robust to the deviation from the within-locus constant selection assumption. However time-independent PRF models, either fixed effect or random effect,

tend to overestimate speciation time. For example, using the exact same data set, Sawyer et al. [43, 44] estimated a divergence time of 4.46 using a fixed effect model and 4.48 by a random effect model. This set of data also contains information with respect to sex bias. Specifically, there are 33 male-biased genes, 28 female-biased genes and 30 sex-unbiased genes (see details in [44]). Similar to Figure 4.3, we generate three plots of estimated selection coefficient respectively for male-biased group, female-biased group and sex-unbiased genes and presented in Figure 4.4.

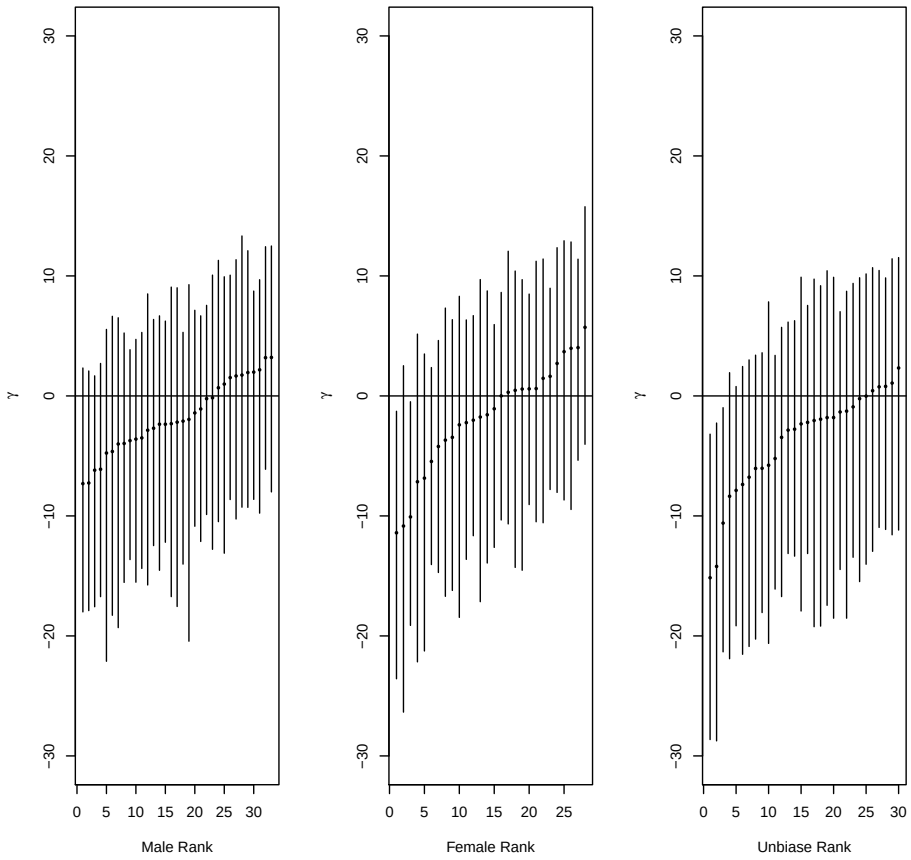

Figure 4.4: Selection parameter γ for male-biased(left), female-biased(middle), and sex-unbiased(right) genes

According to the Figure 4.4, new replacement mutations in sex-biased genes(male- or female-biased) are more likely to be favorable. 30% of male- and 46% of female-biased genes are under positive selection, and male-biased genes have shorter creditable intervals. In contrast, only 16.6% of replacement mutations occurred in sex-unbiased genes are beneficial. It substantially disagrees with the prospect that replacement mutations in sex-biased genes are majorly under adaptive selections[3].

4.3 Estimation of Genetic Proportions

Assuming that within a genetic locus i, each new replacement mutation is subject to a selection coefficient y which is normally distributed with mean γ_i and a variance σ_w. using the 91 genes, we extend our study to infer the following for quantities which are widely applied in the area of population genetics. They are the expected population proportion of beneficial new replacement mutations, the expected proportion of sample polymorphisms due to positive selection, the expected proportion of fixed differences due to positive selection, and finally the mean selection coefficient for fixed differences.

The expected population proportion of beneficial replacement mutations is given by

$$\int_0^{+\infty} N(y|\gamma_i, \sigma_w)dy \tag{4.1}$$

and our estimate gives 0.341. It is coordinated with our estimate of μ_γ since $E(y) = E(E(y|\gamma_i, \sigma_w)) = E(\gamma_i) = \mu_\gamma$. Hence, deleterious replacement mutations arise more frequently than beneficial ones.

The expected proportion of sample polymorphisms due to positive selection is

$$\frac{\int_0^{+\infty} E(O_y + H_y|\gamma_i)N(y|\gamma_i, \sigma_w)dy}{\int_{-\infty}^{+\infty} E(O_y + H_y|\gamma_i)N(y|\gamma_i, \sigma_w)dy} \tag{4.2}$$

and we estimates this proportion as 0.533. Both deleterious and beneficial replacement mutations have nearly equal chance to contribute to sample polymorphisms.

The expected proportion of fixed differences due to positive selection is

$$\frac{\int_0^{+\infty} E(K_p)N(y|\gamma_i,\sigma_w)dy}{\int_{-\infty}^{+\infty} E(K_p)N(y|\gamma_i,\sigma_w)dy} \tag{4.3}$$

and our model gives an estimated of 0.893. Our result is slightly lower than that of Sawyer et al. [44] where they estimated about 95% of fixed differences between the two species are positively selected. The mean selection coefficient for fixed differences is given by

$$\int_{-\infty}^{+\infty} yE(K_p)N(y|\gamma_i,\sigma_w)dy. \tag{4.4}$$

Our estimate of the above quantity is significantly higher than previous studies. It was estimated in order of 10^{-6} in Sawyer et al. [44] and Abel [1] while our estimate has an order of 10^{-5}.

Summary estimates for the above mentioned genetic quantities are presented in Table 4.3 and results are compared among the 33 male-biased, 28 female-biased genes and 30 sex-unbiased genes. In Figure 4.5, we graphically illustrate our estimated of mean proportion of nonsynonymous new mutations that are positively selected (N), mean proportion of sample polymorphisms due to positive selection (S) and mean proportion of fixed differences due to positive selection (F). Again genes are broken down to the three groups with different sex bias. The error bars represent 95% credible intervals.

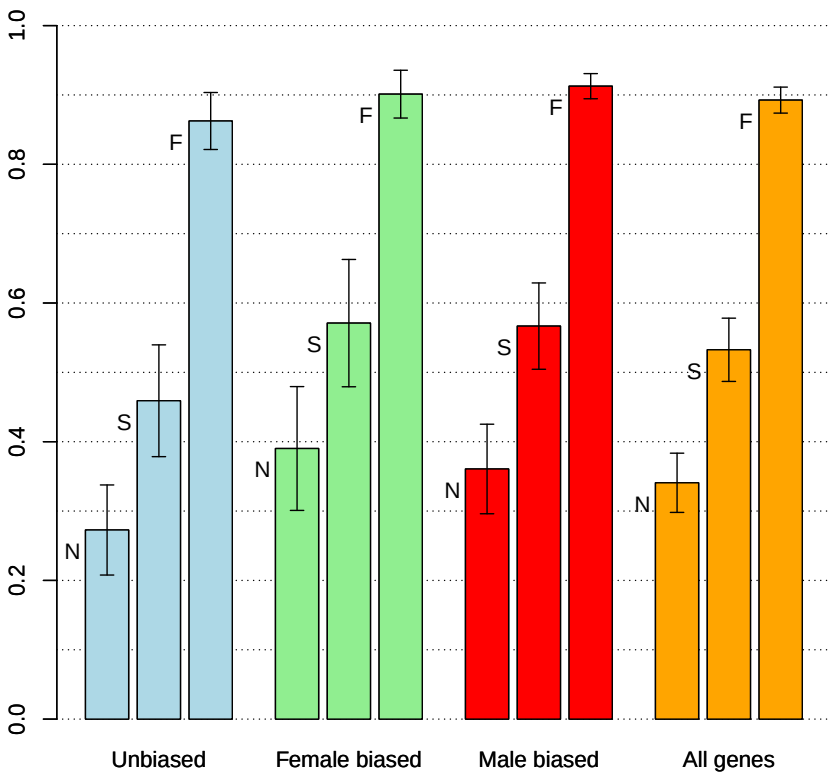

Figure 4.5: Comparison of the three genetic proportions

Feature	Male-biased expression	Female-biased expression	Unbiased expression	All-genes expression
Mean γ of estimated mutational distribution	-2.27	-2.17	-4.34	-2.92
Proportion of new mutations with $N_e s > 0$	0.361	0.39	0.273	0.341
Proportion of sample polymorphisms with $N_e s > 0$	0.567	0.571	0.459	0.533
Proportion of fixed differences with $N_e s > 0$	0.913	0.901	0.862	0.893
Mean $N_e s$ of fixed differences	29.7	39.2	19	29.1

Table 4.3: Comparison of the three genetic proportions

From Table 4.3 and Figure 4.5, sex-biased genes have a greater chance to encounter adaptive selection than unbiased genes. Female-biased genes exhibit slightly stronger positive selection than male-biased genes in both sample polymorphisms and fixed differences. However, the mean γ of fixed differences in female-biased genes is 39.2 which is significantly larger than 29.7 of male-unbiased genes and almost double of the value for sex-unbiased genes. In the Table 4.3, the mutational distributions of γ for male- and female-biased genes are almost identical (-2.27 and -2.17) and the larger mean γ for fixed differences implies that the fixation of replacement mutations in female-biased genes is driven by strong positive selection. The expected proportion of polymorphisms due to positive selection is over 50% and our finding is quite different from Sawyer et al. [44] estimates that beneficial mutations contribute 30% of polymorphisms.

4.4 Conclusion

The time-independent random effect PRF has a tendency to overestimate the divergence time between species[1]. To compensate the overestimation Amei and Sawyer [2], theoretically, developed a time-dependent PRF model by explicitly building the divergence time into the model. Later a time-dependent fixed effect PRF framework with a constant within-locus selection coefficient was applied to aligned DNA sequences of *D.melanogaster* and *D.simulans* to estimate and infer parameters[3]. However, the assumption of constant selection coefficient within a locus is artificial and biologically unrealistic. In the dissertation we relax the constant selection as-

sumption and develop a time-dependent random effect PFR model assuming that, at each locus, each newly-arisen replacement mutation has a selection coefficients y distributed normally with mean γ_i and variance σ_w. In order to make statistical inference about various genetic parameters based on real data set, we applied sample configuration formulas to a hierarchical Bayesian framework. There are two main difficulties in the implementation of the model. One problem is the slow convergence of the underlying Markov chain due to high correlation among parameters and high auto-correlation within certain parameters. Another problem is the extremely long running time of a single iteration due to numerically solving PDEs repeatedly.

We develop a new sampling method called *N-ADAM-Mixing* as well as a parallel computing technique to overcome two main issues. Finally we test our model on simulated date sets and one real data set and results are compared with estimates from previous studies.

Although the model derived in this dissertation is more realistic than previous ones, there are still certain restricted assumptions made in the model. For example, we assume that both species have the same population size. However changes in demographics such as population increase or bottleneck might alter the population size that could shake impact on parameter estimates and confound the interpretation of polymorphism and divergence[6, 15, 17, 28, 32]. Use of *Drosophila* data derived from Africa can avoid some of the demographic complexities[23, 36, 44]. Another assumption is that the nucleotide sites are at high level of recombination and it is equivalent to assuming that sites are at linkage equilibrium. We model nuclotide sites

are independent under this assumption. However, linkage inequilibrium is becoming popular in certain regions of a gene. Further study is needed to check the robustness of the model from departure from these assumptions.

BIBLIOGRAPHY

[1] Abel, H. J. (2009). *The role of positive selection in molecular evolution— Alternative models for within-locus selective effects*. WASHINGTON UNIVERSITY IN ST. LOUIS.

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

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

[4] Bai, Y. (2009). An adaptive directional metropolis-within-gibbs algorithm. *Preprint*.

[5] Bai, Y., Roberts, G. O., and Rosenthal, J. S. (2009). On the containment condition for adaptive markov chain monte carlo algorithms.

[6] Begun, D. J., Holloway, A. K., Stevens, K., Hillier, L. W., Poh, Y.-P., Hahn, M. W., Nista, P. M., Jones, C. D., Kern, A. D., Dewey, C. N., et al. (2007). Population genomics: whole-genome analysis of polymorphism and divergence in drosophila simulans. *PLoS biology*, 5(11):e310.

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

[8] 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(6880):531–534.

[9] Bustamante, C. D., Wakeley, J., Sawyer, S., and Hartl, D. L. (2001). Directional selection and the site-frequency spectrum. *Genetics*, 159(4):1779–1788.

[10] Caccone, A., Amato, G. D., and Powell, J. R. (1988). Rates and patterns of scndna and mtdna divergence within the drosophila melanogaster subgroup. *Genetics*, 118(4):671–683.

[11] Debnath, L. and Mikusiński, P. (2005). *Hilbert spaces with applications*. Academic Press.

[12] Dongarra, J. J., Otto, S. W., Snir, M., and Walker, D. (1995). An introduction to the mpi standard. *Communications of the ACM*.

[13] Dunford, N. and Schwartz, J. T. (1958). Linear operators, vol. i. *Interscience, New York*, 1963.

[14] Ewens, W. J. (2004). *Mathematical population genetics: I. Theoretical introduction*, volume 27. Springer Verlag.

[15] Eyre-Walker, A. (2002). Changing effective population size and the mcdonald-kreitman test. *Genetics*, 162(4):2017–2024.

[16] Fay, J. C., Wyckoff, G. J., and Wu, C.-I. (2001). Positive and negative selection on the human genome. *Genetics*, 158(3):1227–1234.

[17] Fay, J. C., Wyckoff, G. J., and Wu, C.-I. (2002). Testing the neutral theory of molecular evolution with genomic data from drosophila. *Nature*, 415(6875):1024–1026.

[18] Gelman, A., Carlin, J. B., Stern, H. S., and Rubin, D. B. (2004). *Bayesian data analysis*. Chapman & Hall/CRC.

[19] Gelman, A., Roberts, G., and Gilks, W. (1996). Efficient metropolis jumping hules. *Bayesian statistics*, 5:599–608.

[20] Gelman, A. and Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences. *Statistical science*, pages 457–472.

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

[22] Gilks, W. R., Richardson, S., and Spiegelhalter, D. J. (1996). *Markov chain Monte Carlo in practice*, volume 2. Chapman & Hall/CRC.

[23] Glinka, S., Ometto, L., Mousset, S., Stephan, W., and De Lorenzo, D. (2003).

Demography and natural selection have shaped genetic variation in drosophila melanogaster: a multi-locus approach. *Genetics*, 165(3):1269–1278.

[24] Haario, H., Saksman, E., and Tamminen, J. (2001). An adaptive metropolis algorithm. *Bernoulli*, pages 223–242.

[25] Hartl, D. L., Clark, A. G., et al. (1997). *Principles of population genetics*, volume 116. Sinauer associates Sunderland, Massachusetts.

[26] Hastings, W. K. (1970). Monte carlo sampling methods using markov chains and their applications. *Biometrika*, 57(1):97–109.

[27] Itō, K. and McKean Jr, H. P. (1965). *Diffusion Processes and Their Sample Paths: Reprint of the 1974 Edition*, volume 1431. Springer Verlag.

[28] Keightley, P. D. and Eyre-Walker, A. (2007). Joint inference of the distribution of fitness effects of deleterious mutations and population demography based on nucleotide polymorphism frequencies. *Genetics*, 177(4):2251–2261.

[29] Kemeny, J. G., Snell, J. L., and Knapp, A. W. (1966). *Denumerable markov chains*. Van Nostrand Princeton.

[30] Kingman, J. F. C. (1992). *Poisson processes*, volume 3. Clarendon Press.

[31] Lemeunier, F., David, J., Tsacas, L., and Ashburner, M. (1986). The melanogaster species group. *The genetics and biology of Drosophila*, 3:147–256.

[32] McDonald, J. H., Kreitman, M., et al. (1991). Adaptive protein evolution at the adh locus in drosophila. *Nature*, 351(6328):652–654.

[33] Meiklejohn, C. D., Kim, Y., Hartl, D. L., and Parsch, J. (2004). Identification of a locus under complex positive selection in drosophila simulans by haplotype mapping and composite-likelihood estimation. *Genetics*, 168(1):265–279.

[34] Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., and Teller, E. (1953). Equation of state calculations by fast computing machines. *The journal of chemical physics*, 21:1087.

[35] Moran, P. (1959). The survival of a mutant gene under selection. *J. Aust. Math. Soc*, 1:121–126.

[36] Ometto, L., Glinka, S., De Lorenzo, D., and Stephan, W. (2005). Inferring the effects of demography and selection on drosophila melanogaster populations from a chromosome-wide scan of dna variation. *Molecular biology and evolution*, 22(10):2119–2130.

[37] Pröschel, M., Zhang, Z., and Parsch, J. (2006). Widespread adaptive evolution of drosophila genes with sex-biased expression. *Genetics*, 174(2):893–900.

[38] Riesz, F. and Nagy, B. v. S. (1979). Lectures on functional analysis. *Inostrannaya Literatura, Moscow*.

[39] Roberts, G. O. and Rosenthal, J. S. (2004). General state space markov chains and mcmc algorithms. *Probability Surveys*, 1:20–71.

[40] Roberts, G. O. and Rosenthal, J. S. (2007). Coupling and ergodicity of adaptive markov chain monte carlo algorithms. *Journal of applied probability*, pages 458–475.

[41] Sawyer, S. A., Dykhuizen, D. E., and Hartl, D. L. (1987). Confidence interval for the number of selectively neutral amino acid polymorphisms. *Proceedings of the National Academy of Sciences*, 84(17):6225–6228.

[42] Sawyer, S. A. and Hartl, D. L. (1992). Population genetics of polymorphism and divergence. *Genetics*, 132(4):1161–1176.

[43] 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(1):S154–S164.

[44] Sawyer, S. A., Parsch, J., Zhang, Z., and Hartl, D. L. (2007). Prevalence of positive selection among nearly neutral amino acid replacements in drosophila. *Proceedings of the National Academy of Sciences*, 104(16):6504–6510.

[45] Trotter, H. F. (1958). Approximation of semi-groups of operators. *Pacific Journal of Mathematics*, 8(4):887–919.

[46] Venables, W. N., Smith, D. M., and Team, R. D. C. (2002). An introduction to r.

[47] 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. *Proceedings of the National Academy of Sciences*, 102(22):7882–7887.

[48] Yu, H. (2002). Rmpi: Parallel statistical computing in r. *R News*, 2(2):10–14.

[49] Zettl, A. (2005). Sturm-liouville theory, volume 121 of mathematical surveys and monographs. *American Mathematical Society, Providence, RI*.

[50] Zhu, L. and Bustamante, C. D. (2005). A composite-likelihood approach for detecting directional selection from dna sequence data. *Genetics*, 170(3):1411–1421.

VITA

Graduate College
University of Nevada, Las Vegas

Shilei Zhou

Degrees:

Bachelor of Science, Applied Mathematics, 2007
Xidian University, China

Dissertation Title:

Time-Dependent Random Effect Poisson Random Field Model for Polymorphism Within and Between Two Related Species

Dissertation Examination Committee:

Chairperson, Dr. Amei Amei, Ph.D.
Committee Member, Dr. Malwane M.A. Ananda, Ph.D.
Committee Member, Dr. Chih-Hsiang Ho, Ph.D.
Graduate Faculty Representative, Dr. Evangelos A. Yfantis, Ph.D.