EVALUATING THE EFFECTS OF RESCALING PARAMETERS IN LARGE-SCALE GENOMIC SIMULATIONS

A THESIS SUBMITTED TO
THE GRADUATE SCHOOL OF NATURAL AND APPLIED SCIENCES
OF
MIDDLE EAST TECHNICAL UNIVERSITY

BY

OZAN KIRATLI

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR
THE DEGREE OF MASTER OF SCIENCE
IN
BIOLOGY

FEBRUARY 2016

Aproval of the thesis:

EVALUATING THE EFFECTS OF RESCALING PARAMETERS IN LARGE-SCALE GENOMIC SIMULATIONS

submitted by **OZAN KIRATLI** in partial fulfillment of the requirements for the degree of **Master of Science in Biology Department, Middle East Technical University** by,

Prof. Dr. Gülbin Dural Ünver
Dean, Graduate School of **Natural and Applied Sciences** _____

Prof. Dr. Orhan Adalı
Head of Department, **Biology** _____

Assist. Prof. Dr. Ayşegül Ceren Birand Özsoy
Supervisor, **Biology Dept., METU** _____

Examining Committee Members:

Prof. Dr. İnci Togan
Biology Dept., METU _____

Assist. Prof. Dr. Ayşegül Ceren Birand Özsoy
Biology Dept., METU _____

Assoc. Prof. Dr. C. Can Bilgin
Biology Dept., METU _____

Assoc. Prof. Dr. Mehmet Somel
Biology Dept., METU _____

Assoc. Prof. Dr. Ergi Deniz Özsoy
Biology Dept., Hacettepe University _____

Date: 04.02.2016

I hereby declare that all information in this document has been obtained and presented in accordance with academic rules and ethical conduct. I also declare that, as required by these rules and conduct, I have fully cited and referenced all material and results that are not original to this work.

Name, Last name : OZAN KIRATLI

Signature :

ABSTRACT

EVALUATING THE EFFECTS OF RESCALING PARAMETERS IN LARGE-SCALE GENOMIC SIMULATIONS

Kıratlı, Ozan
M.S., Department of Biology
Supervisor : Assist. Prof. Dr. Ayşegül Ceren Birand Özsoy

February 2016, 40 pages

Computer simulations are widely used in many subdisciplines of biological sciences, which evolutionary biology. Large-scale genomic simulations, where several kb (kilo base) to several Mb (megabase) genomes are modeled, are being increasingly used. These simulations require high computing power. There are some methods proposed in the literature to decrease the time and memory demand of these simulations. This study is concentrated on one of those methods, where both the number of generation, and the number of individuals are decreased, and mutation rate is increased. This rescaling method is widely used in recent years. Even though it has been criticized many times, since it could change the population dynamics, there are not many studies that evaluates the effect of rescaling on population dynamics. This study demonstrates how the rescaling of the parameters could change population dynamics with a simple model, and shows that the proportion of polymorphic loci in the simulated genomes could be significantly affected.

Keywords: mutation accumulation, mutation rate, population size, proportion of polymorphic loci, individual based simulations, finite site model, generation time

ÖZ

BÜYÜK ÖLÇEKLİ GENOM SİMÜLASYONLARINDA PARAMETRE YENİDEN ÖLÇEKLENDİRİLMESİNİN ETKİLERİNİN DEĞERLENDİRİLMESİ

Kıratlı, Ozan
Yüksek Lisans, Biyoloji Bölümü
Tez Yöneticisi : Yrd. Doç. Dr. Ayşegül Ceren Birand Özsoy

Şubat 2016, 40 sayfa

Bilgisayar simülasyonları biyolojik bilimlerin, bir tanesi de evrimsel biyoloji olan, bir çok alt dalında yaygın olarak kullanılmaktadır. Birkaç bin bazdan birkaç milyon baza kadar genomik bölgelerin modellendiği büyük ölçekli genom simülasyonlarının kullanımı yaygınlaşmaktadır. Bu simülasyonların hesaplama gücü ihtiyaçları yüksektir. Literatürde bu simülasyonların bellek ve zamansal ihtiyaçlarını azaltmak amacıyla sunulmuş çeşitli yöntemler bulunmaktadır. Bu çalışmada bu yöntemlerden, nesil ve birey sayılarının azaltılarak, mutasyon oranının artırıldığı bir tanesi incelenecektir. Bu yeniden ölçeklendirme yöntemi son yıllarda yaygın olarak kullanılmaktadır. Popülasyon dinamiklerini değiştirebileceği için bir çok kez eleştirilmiş olmasına rağmen, yeniden ölçeklendirme yönteminin popülasyon dinamikleri üzerindeki etkilerini değerlendiren çok fazla çalışma yapılmamıştır. Bu çalışma parametrelerin yeniden ölçeklendirilmesinin popülasyon dinamiklerini nasıl değiştirebileceğini basit bir modelle açıklamakta ve simüle edilmiş genomlardaki polimorfik lokus oranının kayda değer bir şekilde değiştiğini göstermektedir.

Anahtar Kelimeler: mutasyon birikimi, mutasyon oranı, polimorfik lokus oranı, birey bazlı simülasyonlar, sonlu konum modeli, nesil sayısı

To my future cat Schrödinger

ACKNOWLEDGEMENTS

I hereby thank many people who made this impossible thesis possible.

First and foremost, it would not have been possible to write this thesis without the help, support, and patience of my advisor Dr. Ayşegül Birand. Her idealist way of teaching, her mentorship, her continuous advice, and her guidance have taught me to take the first steps in science, all the way I have been progressing throuhout my master's studies.

I would like to thank all my thesis committee members; Prof. Dr. İnci Togan, Assoc. Prof. Dr. C. Can Bilgin, Assoc. Prof. Dr. Mehmet Somel, and Assoc. Prof. Dr. Ergi Deniz Özsoy.

I thank Prof. Dr. İnci Togan, Dr. Erol Akçay, and Dr. Mehmet Somel for being my mentors and role models in the pursuit of being a scientist. I am deeply thankful to my undergraduate advisor, now deceased, the late Prof. Dr. Aykut Kence, for encouraging me to learn and apply as much math as possible to biological sciences, and leading my way, like many other scientists' in the field. I also thank Turkish Scientific and Technical Research Council (TÜBİTAK) for financially supporting me throughout my master's studies.

My family has been a strong anchorage for me that I could count on, they helped me to tackle my obstacles when I needed their encouragement. I thank my father, being a role model for me to overachieve when it is even not possible. I thank my mother for her compassion and wisdom, which will always be the light on my way. I also thank my sister for believing in me throughout my life.

I cannot explain her help in words that I also want to thank Çisel Kemahlı for her friendship, her encouragement, and focusing me to the point when I needed. Her help in the times of crisis was very crucial that this thesis could be submitted.

I want to thank my friends, who supported me throughout my process. Umut Toprak, Mehmet Ali Döke, Emre Erdenk, Mustafa Mısır, Samet Öksüz, Ezgi Dilan, and Ezgi Özkurt have always supported me with their invaluable feedback on my work. I also thank Arev Pelin Sümer, Oğuzhan Beğik, Babür Erdem, Tuğçe Zorlucan, Sevgi Sesli, Eleonora Tafuro, Özgür Can Özüdoğru, Melike Sever, Tammy Tran, and Daniel Barnes for their friendship and their invaluable help.

I also thank many others, who helped me in many cases with their invaluable friendship, apologizing them deeply since their names have not listed here.

Finally, I want to thank to Burcu Duran, for her friendship and encouragement, this thesis would be almost impossible without her support, and encouragement.

TABLE OF CONTENTS

LIST OF TABLES

TABLES

LIST OF FIGURES

FIGURES

LIST OF ABBREVIATIONS

MC	Monte Carlo
MCMC	Markov chain Monte Carlo
MRCA	most recent common ancestor
kb	kilo base pair
Mb	mega base pair
RAM	random access memory
CPU	central processing unit
N	number of individuals in population
μ	mutation rate
t	number of generations
λ	rescaling factor
s	selection coefficient
l	length of genetic material
SNP	single nucleotide polymorphism
Gb	giga base pair
ζ	rate of mutation accumulation in the population
m	total number of mutated loci in the population
ρ	proportion of polymorphic sites
r	proportion of polymorphic loci
ρ_j	proportion of polymorphic loci to be added in a single generation
b	mean number of offspring per individual
b_i	absolute number of offspring of individual i
b_T	total number of offspring in a generation
μ_j	mutation rate for generation j
MWW	Mann-Whitney-Wilcoxon

CHAPTER 1

INTRODUCTION

Ecology and evolutionary biology has taken advantage of mathematics since the beginning of 20th century (Servedio et al., 2014). In these mathematical research in biological sciences, biologists come up with particular equations giving the result of a scenario under specific conditions, and assumptions. These analytical models are powerful; however, they can only be obtained under rather simple scenarios. For more realistic conditions, simulations have become useful tools for biologists to use (Bergstrom et al., 1999; Brown & Rothery, 1993; Hoban et al., 2012; Hudson, 2002; Keen & Spain, 1992). Computer simulations are being used more than ever as high performance computers have become more available recently (García-Dorado & Gallego, 2003; Hoggart et al., 2007; Hudson, 2002; Jiang et al., 2010; Keen & Spain, 1992; Schaffner et al., 2005; Schwartz, 2008). Computer simulations are similar to experiments, where one creates a controlled environment to test various hypotheses (Servedio et al., 2014). Simulations also make it possible to generate a null hypothesis dataset, which can be tested against the empirical data (Carvajal-Rodríguez, 2008; Fusté, 2012; García-Dorado & Gallego, 2003; Hoban et al., 2012; Killcoyne & del Sol, 2014; Kim & Wiehe, 2009; Sargolzaei & Schenkel, 2009).

1.1 Computer Simulations in Biological Research

In biological research, there are numerous types of computer simulations. They can be investigated in terms of their types, time-wise approaches (forward-time or backward-time), and purposes.

1.1.1 Types of Computer Simulations

There are two main types of simulations in biological research: individual (agent) based, and Monte Carlo (MC) simulations (Winsberg, 2015). Individual based simulations are used to simulate many individuals for many numbers of generations, and make it possible to monitor the changes in the population in this long timescale. Typically in individual based simulations, the populations are composed of finite number individuals with certain traits that can mutate, evolve, and selected for or against according to the question or the model itself (Carvajal-Rodríguez, 2008; Hoban et al., 2012; Pickrell et al., 2009; Yuan et al., 2012). Individual based simulations are convenient ways to study populations, since in nature for many organisms, the time-scale to observe the changes in the populations is very long, and also very expensive in the lab, thus such observation is very hard to be achieved in the lifetime of a human. Even with *Escherichia coli*, with such high reproduction rate (i.e. 30 mins per division), it takes many years to observe evolutionary changes. In the very long term evolution experiment by Lenski (2015), 50,000 generations in *E. coli* was achieved in 2010, 22 years after the experiment has started (Lenski, 2011).

MC simulations depend on repeated random samplings and convergence principle, where many samples are generated using random variables, converging to a probability distribution. MC simulations are used when a set of scenarios are expected to happen but the real outcome is not known, or when the outcome is known but the related probabilistic scenario is unknown. They are widely used in Bayesian inference of phylogenies (Duchêne et al., 2015; Gruijter et al., 2011), protein (Jónsson et al., 2012; Paquet & Viktor, 2015; Zhang & Chou, 1992), and membrane studies (Hitsov et al., 2015; Morriss-Andrews & Shea, 2015). In Bayesian inference of phylogenies, a special type of MC methods, Markov chain Monte Carlo (MCMC) simulations are used to generate many phylogenies to decide on the most likely phylogeny (Lanier & Knowles, 2015; Larson-Johnson, 2016; Oaks, 2015). This method, where the similarity scores among many phylogenies are calculated and used with a likelihood function to find the best scenario, is known as the maximum likelihood principle. For protein and membrane studies again the purpose is to generate many possible protein and membrane structures under given thermodynamical conditions (Paquet & Viktor, 2015).

1.1.2 Time-wise Approaches in Computer Simulations

In terms of time, simulations have two main approaches: i.e. forward-time and backward-time simulations (Carvajal-Rodríguez, 2008; Kim & Wiehe, 2009; Tofanelli et al., 2011; Varadarajan et al., 2008; Yuan et al., 2012). Forward-time simulations typically start from an initial state, and go through generations, allowing biologists to monitor the dynamics of interest in the model. Simulations used for speciation research can be given as an example for forward-time simulations (Barraclough & Vogler, 2000; Birand et al., 2012; Gavrilets, 2014). In the simulations used in speciation research, the aim is to understand by which mechanism species originate. These simulations, in this sense, help scientists to generate speciation models that are testable with the empirical data. Backward-time simulations, however, start from a final state (e.g with a population at present), and try to converge to a point back in time (Carvajal-Rodríguez, 2008; Kim & Wiehe, 2009; Yuan et al., 2012). One of the best known backward-time simulations are the coalescent simulations, where the simulation finds the most recent common ancestor (MRCA) of the starting genomes (Carvajal-Rodríguez, 2008; Yuan et al., 2012).

1.1.3 Purposes of Computer Simulations

Biologists could take the advantage of the computer simulations for various purposes (Carvajal-Rodríguez, 2008; Hoban et al., 2012; Tofanelli et al., 2011; Varadarajan et al., 2008). First, biologists may use the computer simulations for predictive purposes, where they test the results from the preexisting mathematical models with simulations (Bergstrom et al., 1999; Brown & Rothery, 1993; Hoban et al., 2012; Hudson, 2002). Speciation research could be grouped under this (Gavrilets et al., 2007; Hoban et al., 2012; Kirkpatrick & Ravigne, 2002). Predictive simulations are also widely used in epidemiology to predict the disease dynamics, and their expansion patterns (Dwyer et al., 1990; Nsoesie et al., 2013).

Alternatively, biologists could use computer simulations to make statistical inferences (Hoban et al., 2012), where biologists generate a distribution of data with simulations, and compare this data against the empirical data to make inferences about the question (White et al., 2014; Winsberg, 2015). For example, evolutionary

histories, and demographic changes can be inferred using these type of simulations, which are used to make statistical inferences (Coop et al., 2009; Enard et al., 2010; Pickrell et al., 2009). MCMC method, which is discussed before, is an example for this type of simulations.

Simulations may also be used for validation of statistical methods, where scientists change the parameters, and try to discover, how the statistical test performs (Enard et al., 2010; Peter et al., 2010). Lastly, simulations may be used for determining the power of the samplings, where simulations are used to understand what kind of sampling is better when designing the research (Meuwissen & Goddard, 2010; Ryman et al., 2006; Spencer et al., 2009).

1.2 Large-scale Genomic Simulations

With the recent advances in computation, the scale of the simulations have reached to a new level, where biologists started to simulate large-scale genomic regions (Carvajal-Rodríguez, 2008). Large-scale genome simulations (Hoggart et al., 2007; Hudson, 2002; Killcoyne & del Sol, 2014; Pickrell et al., 2009; Sargolzaei & Schenkel, 2009; Uricchio & Hernandez, 2014; Varadarajan et al., 2008), which will be covered in detail throughout this chapter, are widely used for generating genomic samples to compare with empirical data. In large-scale genome simulations, typically a genomic region from several thousands (kb) to several million (Mb) base pairs is modeled in an evolutionary time scale (i.e. thousands of generations).

Large-scale genome simulations may adopt forward or backward approaches or even both (Pickrell et al., 2009; Tofanelli et al., 2011; Uricchio & Hernandez, 2014; Varadarajan et al., 2008; Yuan et al., 2012). In this study, we will concentrate on a specific type of large-scale genome simulations, namely forward-time individual based large-scale genomic simulations. These simulations, typically assume finite sized genomes since each locus is modeled one by one. In population genetics, however the assumption of infinite genome size is not rare (Griffiths, 1982; Hudson, 1991; Kimura & Crow, 1964; Kimura & Ohta, 1969; Kimura, Ohta, & MacArthur, 1971; Ma et al., 2008; Tajima, 1996; Watterson, 1975). Since, the genome size of many organisms are relatively high, infinite size genome assumes infinitely many loci, where every mutation causes a new allele in the population

and the probability of back mutation is negligibly low. This assumption simplifies the calculations of the expectation values of homozygosity and the number of alleles present in the population.

1.2.1 Limitations of Large-scale Genomic Simulations

Biological simulations are limited by the ultimate problems of computing; running time, and memory demand (Hoggart et al., 2007; Keen & Spain, 1992; Schaffner et al., 2005). Large-scale genomic simulations in particular have high time and memory demands (Carvajal-Rodríguez, 2008; Hoggart et al., 2007; Schaffner et al., 2005; Thornton, 2014). Mainly, there are two obstacles causing this demand. First, the computational power of any computer is limited (Branke, Kaussler, Smidt, & Schmeck, 2000). That is to say, if one wants to simulate whole human genome, which is almost 2900Mb (International Human Genome Sequencing Consortium, 2004), it is still unlikely to run a simulation for this size of genome, even with a very few individuals (e.g. 5), and for a few generations (i.e. 10), because of the very high random-access memory (RAM) and central processing unit (CPU) need. Today, the simulations of 20Mb genomic regions are considered as large-scale genomic simulations.

Second limitation arises since each coding language and/or algorithm have their own limitations (Gen & Cheng, 2000). Each algorithm's CPU and RAM need will determine the limits of that algorithm, that some of them will run faster and/or use less memory than the others. In addition to the algorithms, the coding language also changes the memory and time need dramatically. In a lower level coding language, with less built-in functions, and user interfaces, the CPU and RAM demands are not going to be high, however coding with such language will be challenging for the developer (García-Dorado & Gallego, 2003; Hoban et al., 2012).

1.3 Efforts to Increase Time and Memory Efficiency

There are numerous efforts to make the simulations less time and memory consuming (Chadeau-Hyam et al., 2008; Hoggart et al., 2007; Pickrell et al., 2009; Ruths & Nakhleh, 2013; Sargolzaei & Schenkel, 2009; Thornton, 2014). One of the approaches is to change the algorithm, where a more efficient solution to a prob-

lem is adopted and applied to code a new simulation (Killcoyne & del Sol, 2014; Ruths & Nakhleh, 2013; Thornton, 2014). For example, Ruths and Nakhleh (2013) suggested an algorithmic method in their simulation, where only the mutant loci in the genome are processed instead of all loci to increase time and memory efficiency.

The second approach is to change the parameters in the simulation, with the assumption that the results will be unchanged (Hoggart et al., 2007; Kim & Wiehe, 2009; Yuan et al., 2012). Since there are a finite number of individuals and generations in an individual based simulation, if one decreases the number of individuals and/or generations in the simulation, the total time of the simulation will decrease. Decreasing number of individuals will also decrease the memory demand. Decreasing the generation time will similarly decrease the number of total iterations, and eventually the processing time. In the next section, we will describe this rescaling approach by Hoggart et al. (2007), which is the topic we investigate in this thesis.

1.3.1 Hoggart's Rescaling Approach in Large-scale Genomic Simulations

Hoggart et al. (2007), and Chadeau-Hyam et al. (2008) proposed a rescaling method, in which they rescaled the population size (N), mutation rate (μ), and number of generations (t) to increase the time and the memory efficiency. In this rescaling approach, t and N were decreased by a rescaling factor (λ), whereas μ was increased by the same factor. Consequently, the running time was decreased by λ^2, and memory consumption was decreased by λ. In their study, Hoggart et al. (2007) used $\lambda = 10$ (see Table 1 in Hoggart et al., 2007), and claimed that the average value of number of mutations and heterozygosity did not change after rescaling (see Equation 2 and Fig. 2 in Hoggart et al., 2007). The average number of mutations is often referred as the rate of mutation accumulation (Pickrell et al., 2009), which is the ratio of number of mutations to the size of the genetic material. We will later refer to this method as "Hoggart's rescaling method" (2007).

He et al. (2012) claimed that this approach might keep the values "mutation rate per locus per generation" ($2N\mu$, where μ is mutation rate per locus per gamete per generation), and "selection coefficient per locus per generation" ($4Ns$, where s is selection coefficient per locus per gamete per generation) unchanged after rescaling, since these values are not for a particular locus for one individual but for all pop-

ulation in one generation. According to these results (He et al., 2012; Hoggart et al., 2007), a rescaling of these parameters seems to be plausible for decreasing time and memory demand, while keeping genetic makeup of the population unchanged.

1.3.2 Applications of Rescaling Approach

Hoggart et al. (2007), and Chadeau-Hyam et al. (2008)'s software FREGENE, and their rescaling approach are widely used in the literature. Numerous researchers used FREGENE to generate Single Nucleotide Polymorphism (SNP) data for different purposes with a very wide range of parameters, where N ranges between 500 and $20,000$, and genome length (l) between 70 kb and 20 Mb (Table 1.1).

Many other scientists adopted the rescaling method by Hoggart et al. (2007), while running their own simulations (Coop et al., 2009; Duque et al., 2014; Duque & Sinha, 2015; Gruijter et al., 2011; He et al., 2012; Lohmueller, 2014; Lohmueller et al., 2011; MacLeod et al., 2009; Peng & Amos, 2010; Pickrell et al., 2009; Wu et al., 2009). Coop et al. (2009), and Pickrell et al. (2009) used rescaling method by Hoggart et al. (2007), while running their simulations in *"cosi"* (Schaffner et al., 2005), where they generate SNP data for human populations, and compare it with HapMap (The International HapMap Consortium, 2005, 2007) data to account for recent adaptations in human populations. MacLeod et al. (2009) used the rescaling method in their simulations to generate SNP data to develop a linkage disequilibrium estimator for quantitative trait loci mapping studies. Wu et al. (2009) used the rescaling method in their simulations to detect deletions.

Peng and Amos (2010) designed their own simulation for generating human genome like samples using the rescaling method developed by Hoggart et al. (2007). Gruijter et al. (2011), and He et al. (2012) used the same method in their simulations to generate SNP data to use in positive selection studies. Moreover, rescaling was used by Lohmueller et al. (2011), and Lohmueller (2014) with another simulation, SFS_CODE, to generate SNP data to use in directional selection studies. Finally, Duque et al. (2014), and Duque and Sinha (2015) used it with another simulation PEBCRES, to test the power of the molecular clock models. As a final remark, He et al. (2012), and Duque and Sinha (2015) applied rescaling with a very high rescaling parameter, $\lambda = 1000$, while the general practice is to use $\lambda = 2, 5,$ or 10.

Table 1.1: Parameters used in the studies that adopted FREGENE (Chadeau-Hyam et al., 2008; Hoggart et al., 2007)

Publication	N	t	l	**Purpose**
Ding et al. (2008)	500	$20,000$	50-200 kb	Building phylogenies
López Herráez et al. (2009)	*NA*	*NA*	2 Mb	Analysis for recent positive selection in human populations
Tachmazidou et al. (2008)	$20,000$	*NA*	≥ 1 Mb	Analysis of SNPs with a MCMC method for comparison of different clustering methods
Powell et al. (2010)	1000	$100,000$	50 Mb	To calculate identity by descent
Ayers and Cordell (2010)	$10,000$	*NA*	20 Mb	To test the performance of different logistic regression models
Vounou et al. (2010)	$10,000$	$200,000$	20 Mb	To generate data to compare with empirical data in a neuro-imaging study
Enard et al. (2010)	$25,000$	$150,000$	6 Gb *	To determine the loci subjected to positive selection among primates
Tachmazidou et al. (2010)	$10,000$	*NA*	1 Mb	To test validity of the statistical tests
Tachmazidou et al. (2011)	$20,000$	*NA*	5 Mb	To test the power of statistical tests
Cule and De Iorio (2012) and Cule and De Iorio (2013)	$21,000$	*NA*	$10,000$	To test a statistical method to choose ridge parameter
Yan et al. (2014)	1000	$20,000$	$10,000$	To suggest a method to increase speed of neuro-imaging genetics analysis

* Total genomic size of population.
NA: data not available

1.3.3 Criticism of Rescaling Approach

Although the rescaling method has been used by many researchers, it has also been criticized by many since it might underestimate some "unrecognized" population genetic effects that depend on the absolute values of the population size, number of generations, and the mutation rate (Kim & Wiehe, 2009; Peng & Amos, 2010; Ruths & Nakhleh, 2013; Sargolzaei & Schenkel, 2009; Tofanelli et al., 2011). Hoggart et al. (2007) acknowledged that back mutations, when become more frequent, may affect estimation of recombination rates, and suggested that removing some of the double hit loci would solve the problem.

Ruths and Nakhleh (2013), and Sargolzaei and Schenkel (2009) argued that this rescaling method did not completely solve the time and memory efficiency problems, instead they proposed algorithmic solutions, where they did not process all the genomic region as explained above. Kim and Wiehe (2009), in their review, claimed that the number of individuals may change Tajima's D, which is highly dependent on the population size. Tofanelli et al. (2011) pointed out the "*de facto*" sampling error introduced by rescaling the population size may in turn increase the effect of random genetic drift. Peng and Amos (2010) stated that this rescaling method cannot be used when non additive effects are present, since this approach uses the assumptions of diffusion approximation, which hold only under weak genetic effects. Gavrilets (2005) also addressed the problem of using high mutation rates in simulations, which could erase the effect of random genetic drift.

1.4 Objectives

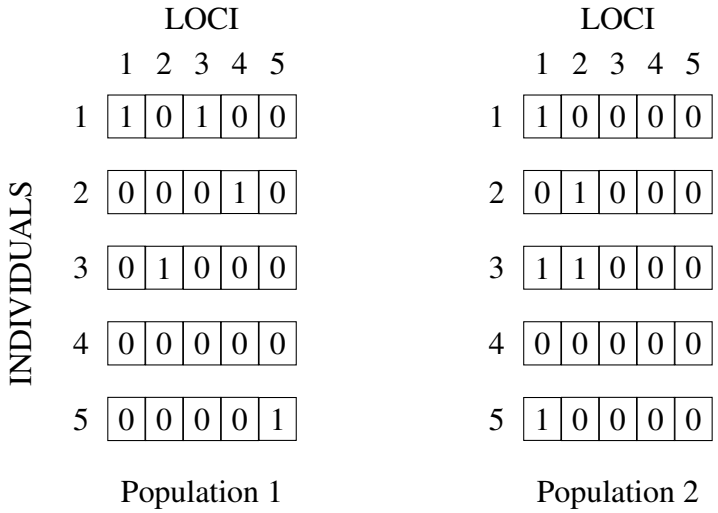
Hoggart et al. (2007)'s rescaling method is widely used for large-scale genome simulations, however, an important but overlooked effect of this rescaling approach could be on the random genetic drift as addressed by Gavrilets (2005), and/or on the proportion of polymorphic loci, which is the rate of the loci having more than one allele to the total number. To the best of my knowledge, these have not yet been evaluated. The problem with the proportion of polymorphic loci might arise since the model used in the large-scale genomic simulations is a finite site model, yet making infinite site model assumptions; the problem with genetic drift might arise from the mutation-drift balance.

We hypothesize that the reason for the proportion of polymorphic loci to be a problem could be because the rate of mutation accumulation, and the proportion of polymorphic loci correspond to different dynamics in populations. Let's consider two populations, each with 5 individuals each having a genome consisting 5 loci, and let ancestral state be "0", and mutated state be "1". Figure 1.1a shows a case, where the total numbers of mutations are the same for two populations whereas the numbers of polymorphic loci are different. Figure 1.1b shows a case where the numbers polymorphic loci are equal, and the total numbers of mutations are different. It is clear that the populations 1, and 2 in the figure have different genetic makeups.

The rate of mutation accumulation in a population is important since it is the source of the variation, and it might be related to other dynamics; such as extinction (Higgins & Lynch, 2001; Lynch et al., 1995), parasitism (Howard & Lively, 1994), and the evolution of sexual reproduction (Muller, 1932, 1964). Number of polymorphic loci, on the other hand, is different than mutation accumulations, since it changes the mutational variation in the population by adding alleles to ancestral loci. This may seem unimportant when all loci are neutral; however, the problem could become significant when the loci are not neutral.

My aim is to understand the dynamics of the populations when the parameters are rescaled. To understand the dynamics of mutation accumulation, proportion of polymorphic loci, and random genetic drift, we developed an individual based simulation model to answer three main questions: 1) Can we use the rescaling method proposed by Hoggart et al. (2007) without changing the expected genetic makeup of the population? 2) Is proportion of polymorphic loci affected from rescaling of the parameters? 3) Can we develop an analytical solution that explains the behavior of mutation accumulation, and proportion of polymorphic loci?

a)



b)

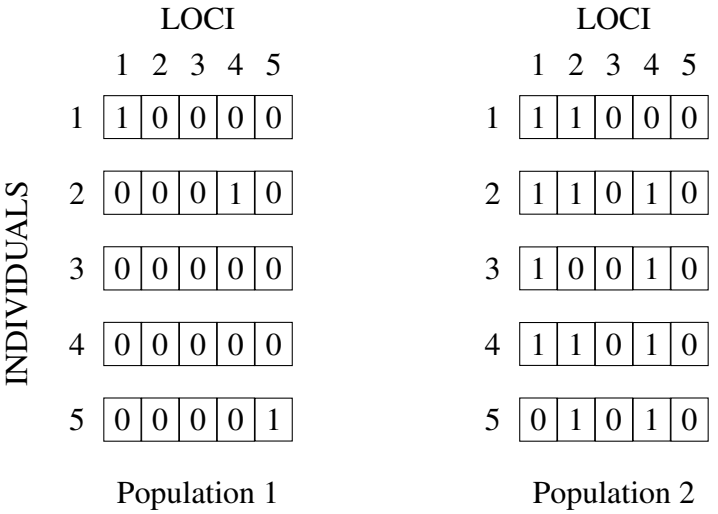


Figure 1.1: Hypothetical scenario demonstrating how the dynamics of rate of mutation accumulation and proportion of polymorphic loci could be different. Each row represents an individual, and each column a locus, where the state 0 represents ancestral state and the state 1 represents mutated state. a) Both populations on the left and right have the 5 mutations, yet the number of polymorphic loci in population 1 is 5, and in the population 2, it is 2. b) Both populations 1 and 2 have 3 polymorphic loci, yet the number of mutations in the population 1 is 3, and in population 2, it is 12.

CHAPTER 2

THEORETICAL CALCULATIONS

In this chapter, we will demonstrate how rate of mutation accumulation, and the proportion of polymorphic loci would be affected by rescaling of parameters.

2.1 Rescaling Method for Keeping Rate of Mutation Accumulation Unchanged

Let the ratio of mutation accumulation be ζ, which is the ratio of total number of mutated loci in the population (m) to the total number of loci in the population consisting of N individuals with l number of loci:

$$\zeta = \frac{m}{Nl} \tag{2.1}$$

If in this population, the mutation rate is equal to μ, and there are t generations, the probability of one locus to be mutated at the end of t generations is μt. If the population is haploid, there will be Nl loci in the population, then m will be $\mu t Nl$, and equation 2.1 becomes;

$$\zeta = \frac{\mu t Nl}{Nl} = \mu t \ . \tag{2.2}$$

The above equation shows that mutation accumulation (ζ) does not seem to have a dependency on N, it is only related to μ and t. It also does not matter whether the population is haploid or diploid (instead of N, $2N$ will be both in the numerator and denominator, and will cancel out).

Therefore, to keep the ζ constant, a rescaling on μ and t can be applied. Let, μ'

be rescaled mutation rate by a rescaling parameter, λ, that is $\mu' = \mu\lambda$. Similarly, let t' be the rescaled number of generations by the same factor λ, which is $t' = t/\lambda$. Then, ζ will still remain constant, since;

$$\mu' t' = \mu\lambda \frac{t}{\lambda} = \mu t = \zeta \, . \tag{2.3}$$

We will refer to this method as "rescaling for mutation accumulation", where only μt are rescaled, and N is not. It is different from Hoggart's rescaling method (2007) in which N is also rescaled by λ.

2.2 Rescaling Method for Keeping Proportion of Polymorphic Loci Unchanged

Hoggart et al. (2007) assume that the mutations are rare, and there are many loci. Keeping these assumptions in mind, we will concentrate on the proportion of polymorphic loci. Let the proportion of polymorphic loci (ρ) be the ratio of total number of polymorphic loci (r) to the size of the genome:

$$\rho = \frac{r}{l} \, . \tag{2.4}$$

Here, calculating r is not as straight-forward as calculating m. Again, as mentioned above, the probability of a mutation in one particular locus in a single generation is μ, but this time, we are interested in the probability of finding at least one different allele on one locus among all the individuals in the population. If there were infinite number of loci in the genome, in a single generation, and if mutations are very rare, then the probability of a mutation to occur in the exact same locus in different individuals and/or in the same individual (i.e. back mutations) would be very low.

Making the above-mentioned assumptions, in a single generation, it is expected that there would be $N\mu l$ mutations. Let the expected proportion of polymorphic loci to be added in a single generation be ρ_j, where j represents a single generation (i.e. $j \in \{1, \ldots, t\}$). Then:

$$\rho_j = \frac{N\mu l}{l} = N\mu \, , \tag{2.5}$$

since there are l number of loci. This suggests that $N\mu$ amount of all loci will become polymorphic at each generation. Note that, this can only be true where $N\mu \ll 1$, since as this ratio will get closer to 1, the probability that two or more mutations to occur at the same locus increases, which will eventually decrease ρ_j, and make it, $\rho_j < N\mu$. Also, the equation 2.5 is only applicable for models with infinite loci because no matter how much mutation occurs in each generation, this will not cause an accumulation of polymorphic loci in the genome. Mutations will continue to appear in different loci, so in each generation $N\mu$ proportion of new loci will continue to add. In equation 2.5, it is shown that the proportion of polymorphic loci to be added in each generation is $\rho_j = N\mu$, and since there are t number of generations:

$$\rho = \rho_j t = N\mu t \, . \tag{2.6}$$

From this equation it is clear that a rescaling as mentioned by Hoggart et al. (2008) cannot be done since;

$$N'\mu't' = \frac{N}{\lambda}\mu\lambda\frac{t}{\lambda} = \frac{N\mu t}{\lambda} \, , \tag{2.7}$$

therefore,

$$N\mu t \neq N'\mu't' \, . \tag{2.8}$$

If all above mentioned assumptions were true, according to the equation 2.6, we would expect a linear increase of number of polymorphic loci by time. With the previously stated criticisms in mind, we want to remind that these large-scale genome simulations, while making infinite site assumptions, use finite site models. This suggests that in these large-scale genome simulations with increasing values of $N\mu$, and as the generation number increases in a single simulation, the probability of having new mutations at the same locus of different individuals increase. This will make ρ_j less than $N\mu$ for a single generation, making ρ less than $N\mu t$, which would complicate this dynamic further. Since, the expectation for ρ cannot exceed 1, it can be roughly summarized as;

$$N\mu \leq \rho \leq \min(1, N\mu t) \, . \tag{2.9}$$

From the above solutions, for ρ, it can be said that it is primarily related with $N\mu$ (Equation 2.5), and it also accumulates with time. Hence, when t is rescaled, a rescaling of N and μ is not enough to solve the problems that will arise from the

proportion of polymorphic loci. However, a rescaling for keeping only the proportion of polymorphic loci unchanged, would be possible between N and μ, only when keeping t constant (Equations 2.5 and 2.6). We will refer to this type of rescaling as "rescaling for polymorphic loci".

CHAPTER 3

MODEL AND SIMULATIONS

In this chapter, we lay out the details of the individual based simulation model that we developed, to test the effects of Hoggart's rescaling method (2007) on rate of mutation accumulation, and proportion of polymorphic loci.

3.1 Assumptions

We assume that the population is haploid, the generations are non-overlapping, the population size is constant, and the reproduction is asexual. All loci are biallelic, and all mutations are only point mutations, and back mutations are allowed. There are no insertions, deletions, or inversions. We also assume that there are no selection, no migration, and no recombination.

3.2 Population

The population is composed of individuals, which have binary sequences that represent their genomes with l number of loci. In these sequences, 0 represents ancestral state, and 1 represents mutated state. Initially, a matrix of N individuals with l loci, is produced, where each row represents a genome of an individual, and each column represents a particular loci. The simulation starts with identical individuals having all zeros (ancestral state). As there are N individuals, and population size is constant, the population matrix size does not change throughout the simulation.

3.3 Life Cycle

Individuals go through a simple life cycle where they produce offspring, and they die. Some of the offspring produced become adults in the next generation, and the life cycle is repeated again.

3.3.1 Offspring Generation

The number of offspring for each individual is assigned randomly with a Poisson distribution around a mean b (Gillespie, 1975). In each generation an array of Poisson distributed random numbers is generated, representing their number of offspring ($b_i, i = 1, .., N$). First number, b_1, is applied to first individual (the first row of population matrix). So, the first row of the population matrix is copied to a new offspring matrix, b_1 times. Second individual is copied b_2 times, and same procedure applied for all N individuals in population matrix. Total number of offspring (b_T) changes every generation.

3.3.2 Mutations

Mutation rate (μ) is the probability of mutations per locus per gamete per generation. In order to introduce this stochastic nature of mutations in the simulations, we draw a random value for the actual mutation rate from a normal distribution of mutation rates with a mean μ and a standard deviation of 0.1μ. Actual mutation rates for each generation are generated at the beginning of the simulation (μ_j, $j = 1, ..., t$). With this rate, for each generation a mutator matrix with same size as offspring matrix ($b_T \times l$) is generated. The mutation values (i.e. 1s) are randomly distributed to matrix with a μ_j ratio.

3.3.3 Choosing New Parents

Since the population size is assumed to be constant, only N number of the offspring survives. In each generation, N number of random and non-repeating integers between 1 and b_T are generated. The rows in the offspring matrix corresponding to the generated numbers are copied and replaced in the population matrix. This is done for each generation, and the population is replaced with surviving offspring at the end of each generation. Note that, since the population size is assumed to be constant, the simulations where $b_T < N$ are aborted.

3.4 Parameters

The parameters we change during the simulations are N, μ, and t, which are given below (see Table 3.1). The range of parameter values are chosen mostly similar to those used in literature (see Table 1.1) with $\lambda = 10$. The parameters, which were not changed during the simulations were, $l = 10000$ and $b = 2$. Results presented in the next chapter are based 10 runs for each parameter set.

Table 3.1: The parameters used in simulations

	N	μ	t
Set 1	1000	10^{-5}	1000
Set 2	10000	10^{-6}	10000
Set 3	1000	10^{-5}	10000
Set 4	1000	10^{-6}	10000
Set 5	10000	10^{-5}	1000
Set 6	10000	10^{-6}	1000

To evaluate Hoggart's rescaling method (2007), we will compare sets 1 and 2 (comparison A, Table 3.2). To evaluate whether rate of mutation accumulation change according to calculations we demonstrated in chapter 2, we will compare sets 1 vs 4, and sets 2 vs 5 (comparisons B and C respectively, Table 3.2). Similarly, to evaluate the calculations on rescaling for polymorphic loci, we will compare sets 1 vs 6 and sets 2 vs 3 (comparisons D and E respectively, Table 3.2).

3.5 Theoretical Expectations

Based on the calculations discussed in chapter 2, we have four main expectations. First, in comparison A, despite Hoggart et al. (2007)'s claim that the genetic make up of the population will not be altered with "Hoggart's rescaling method", we expect to find significant difference for proportion of polymorphic loci, but no significant difference for mutation accumulation. The latter was only the tested dynamics

Table 3.2: List of comparisons with the parameters used

Comparison	Sets	N	μ	t	Method	Rescaled Parameters
Comparison A	Set 1*	1000	10^{-5}	1000	Hoggart's rescaling	$N,\ \mu,\ t$
	Set 2	10000	10^{-6}	10000		
Comparison B	Set 1*	1000	10^{-5}	1000	Rescaling for mutation accumulation	$\mu,\ t$
	Set 4	1000	10^{-6}	10000		
Comparison C	Set 2	10000	10^{-6}	10000	Rescaling for mutation accumulation	$\mu,\ t$
	Set 5*	10000	10^{-5}	1000		
Comparison D	Set 1*	1000	10^{-5}	1000	Rescaling for polymorphic loci	$N,\ \mu$
	Set 6	10000	10^{-6}	1000		
Comparison E	Set 2	10000	10^{-6}	10000	Rescaling for polymorphic loci	$N,\ \mu$
	Set 3*	1000	10^{-5}	10000		

* The marked sets are the rescaled sets.

in Hoggart et al. (2007). Second, similarly for comparisons B and C, where we evaluate the "rescaling for mutation accumulation" method, we expect no significant difference for mutation accumulation, since $\mu t = 10^{-2}$ in all the sets, and significant difference for the proportion of polymorphic loci, since at least one of $N\mu$ or t is different in these comparisons. Third, the proportion of polymorphic loci will show no significant difference for the comparisons D and E, where we evaluate the "rescaling for polymorphic loci" method, since both $N\mu$ and t values are the same in those sets ($N\mu = 10^{-1}$, $t = 1000$ for comparison D, and $N\mu = 10^{-1}$, $t = 100$ for comparison E); however, the rate of mutation accumulation will be significantly different, since μt values of pairs are different. Forth, based on my calculations in section 2.1, we expect that the rates of mutation accumulation for all four sets 1, 2, 4, and 5 to show no significant difference, since mutation accumulation depends only on μt but not N.

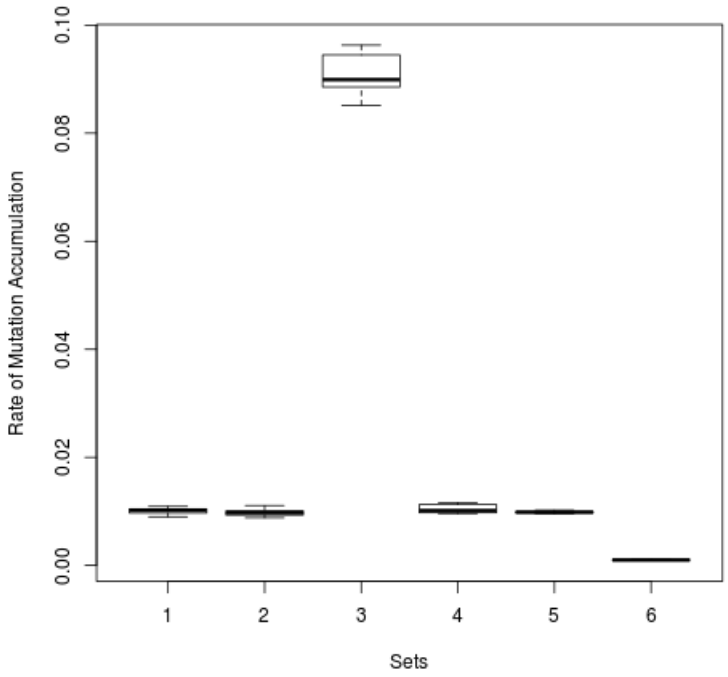
CHAPTER 4

RESULTS

All statistical tests are done using R (R Core Team, 2015). We compared pairs of sets, in terms of rate of mutation accumulation and proportion of polymorphic loci, and since we have a few samples for each group, we applied Mann-Whitney-Wilcoxon (MWW) test to test if their medians are equal or not. The results of the simulations based on 10 runs for each parameter set are presented in figure 4.1 .

In order to evaluate the effects of Hoggart's rescaling method (Hoggart et al., 2007), and compare that with my expectations based on the calculations laid out in chapter 2, we will present the results as laid out in table 3.2. The rate of mutation accumulation is not significantly different in comparison A (Hoggart's rescaling method), however, the medians of the proportion of polymorphic loci are significantly different (Table 4.1 & Fig. 4.2), which is in agreement with the first expectation we laid out in section 3.5. Similarly for comparisons B and C (rescaling for mutation accumulation), there is no significant difference between the medians for mutation accumulation, however, the medians for proportion of polymorphic loci are significantly different (Table 4.1 & Fig. 4.2), which is consistent with the second expectation. The difference between the medians of rate of mutation accumulation are significant for comparisons D and E (rescaling for polymorphic loci), however, the difference between the medians for number of polymorphic loci are not significant (Table 4.1 & Fig. 4.3), which are also consistent with the third expectation.

a)



b)

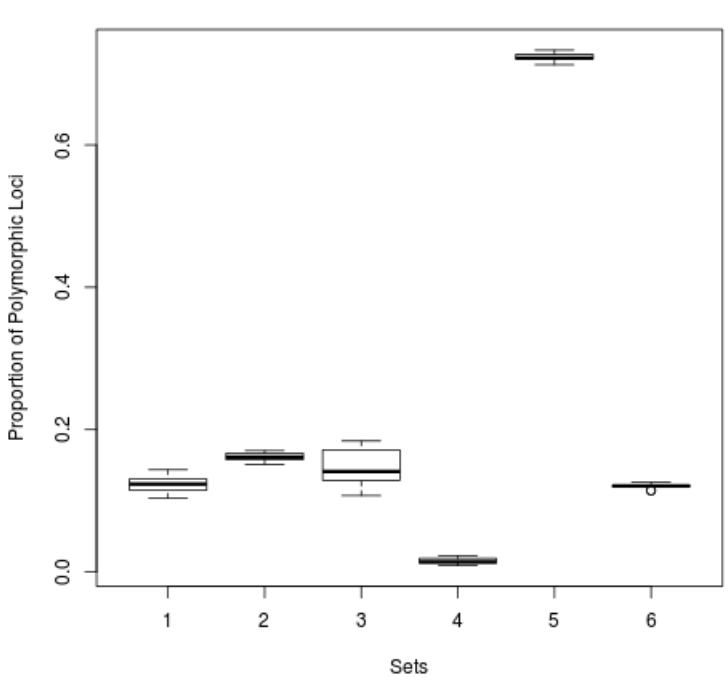


Figure 4.1: a) Rates of mutation accumulation, b) proportions of polymorphic sites of the sets given in Table 3.1

Table 4.1: p values for comparisons from MWW tests for the rate of mutation accumulation and the proportion of polymorphic loci in the simulations. Bold values represents significant difference in 95% confidence interval

Comparison	Sets	N	μ	t	Rate of Mutation Accumulation	Proportion of Polymorphic Loci
Comparison A	Set 1*	1000	10^{-5}	1000	3.53×10^{-1}	$\mathbf{1.08 \times 10^{-5}}$
	Set 2	10000	10^{-6}	10000		
Comparison B	Set 1*	1000	10^{-5}	1000	4.36×10^{-1}	$\mathbf{1.08 \times 10^{-5}}$
	Set 4	1000	10^{-6}	10000		
Comparison C	Set 2	10000	10^{-6}	10000	5.79×10^{-1}	$\mathbf{1.08 \times 10^{-5}}$
	Set 5*	10000	10^{-5}	1000		
Comparison D	Set 1*	1000	10^{-5}	1000	$\mathbf{1.08 \times 10^{-5}}$	5.96×10^{-1}
	Set 6	10000	10^{-6}	1000		
Comparison E	Set 2	10000	10^{-6}	10000	$\mathbf{1.08 \times 10^{-5}}$	1.43×10^{-1}
	Set 3*	1000	10^{-5}	10000		

In the fourth expectation, the values of rates of mutation accumulation of sets 1, 2, 4, and 5 are expected to show no significant difference, since we expect it to have no dependency on N. The results are as expected, since p values of pairwise comparisons between sets 1 vs 5 ($p = 0.31$), sets 2 vs 4 ($p = 0.07$), and sets 4 vs 5 ($p = 0.19$) showed that the difference between these sets are non-significant. We also applied Kruskal Wallis test for multi group comparisons, and the difference between these four sets is not significant ($p = 0.27$). Thus, both pairwise comparisons using MWW tests, and Kruskal Wallis showed that the theoretical expectations are consistent with the results.

The rate of mutation accumulation is the highest in set 3 (Fig. 4.1), and the proportion of polymorphic loci is the highest for set 5 (Fig. 4.1), which are expected results according to the calculations. Mutation accumulation is directly related to μt value (see equation 2.2), and it is the highest for set 3 (i.e. $\mu t = 0.1$). Proportion of polymorphic loci is related to $N\mu$ and t, set 5 has the highest $N\mu$ value ($N\mu = 0.1$).

One other unevaluated expectation that was mentioned in the literature was related to the dynamics of random genetic drift when the parameters are rescaled. Figure 4.4 presents the proportion of fixed loci in the sets simulated, and it can be seen that only in the simulations with $10,000$ generations (i.e. sets 2, 3, and 4; see Table 3.1) there were some fixations (with the exception of 1 run of set 1, which is seen as a single dot in Fig. 4.4). Set 3 had the highest proportion of fixed loci (Fig. 4.4), again due to the highest μt value used, but not due to actual mutation rates that were assigned randomly in the simulations (Fig. 4.5). Finally, set 3 had the highest variance among all sets for all values (Figs. 4.1 and 4.4), which is again not due to the actual mutation rate (Figs. 4.5).

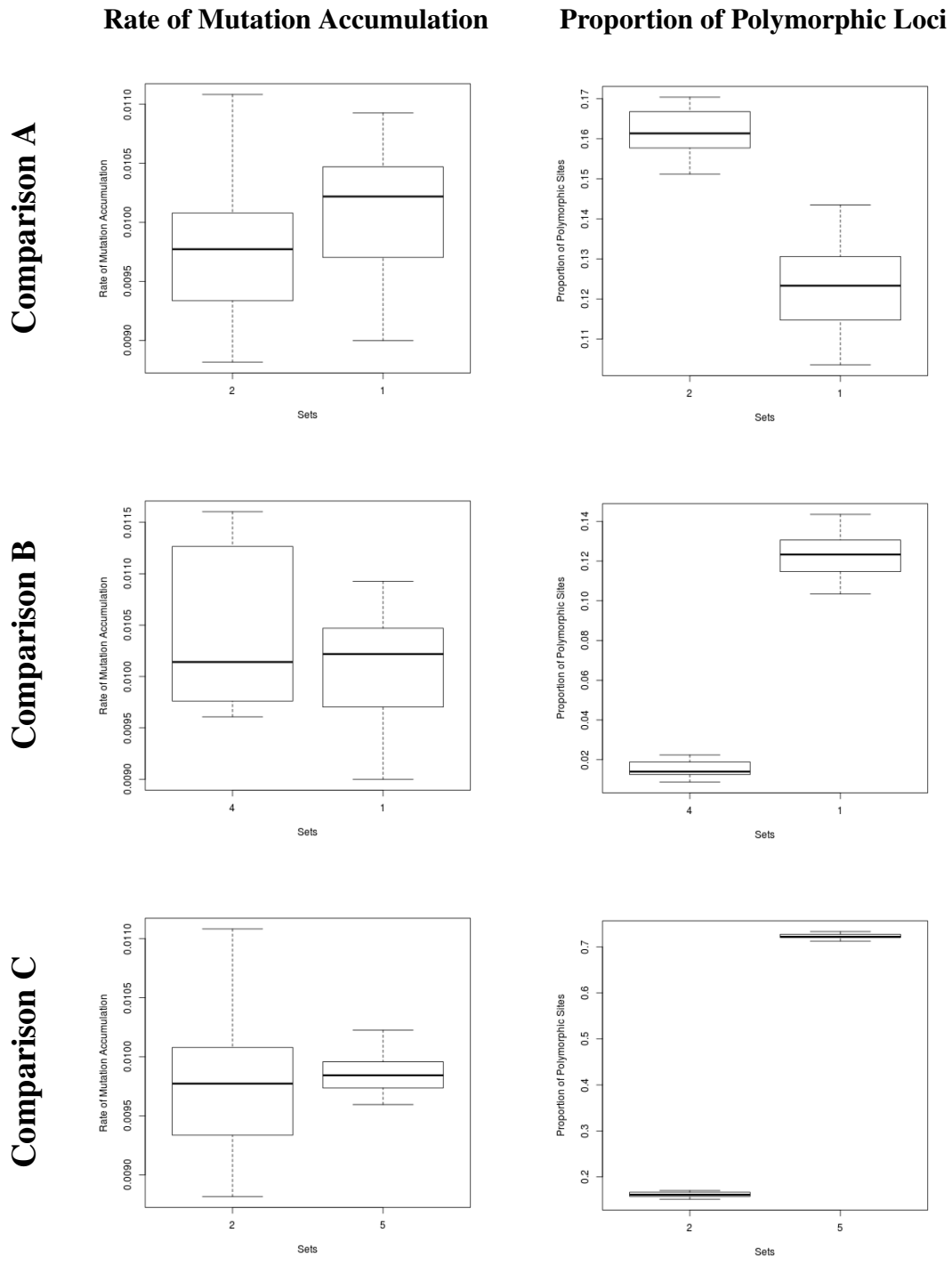


Figure 4.2: Boxplots of comparisons A, B, and C for rate of mutation accumulation (left) and the proportion of polymorphic loci (right). In all boxplots, rescaled sets are on the right of each boxplot. The rates of mutation accumulation are not significantly different in comparisons A, B, and C, but, the proportions of polymorphic sites are.

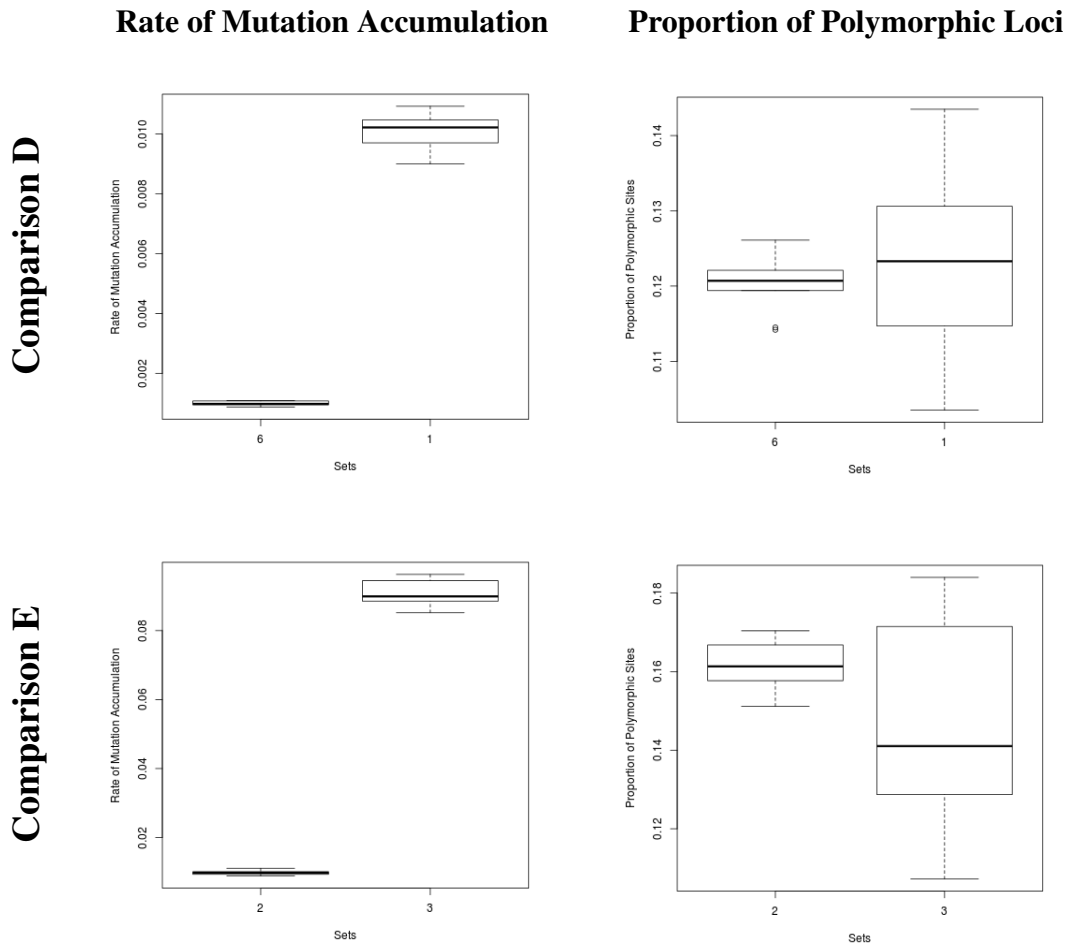


Figure 4.3: Boxplots of comparisons D and E for rate of mutation accumulation (left) and the proportion of polymorphic loci (right). In all boxplots, rescaled sets are on the right of each boxplot. The proportions of polymorphic sites are not significantly different in comparisons D and E, but, the rates of mutation accumulation are.

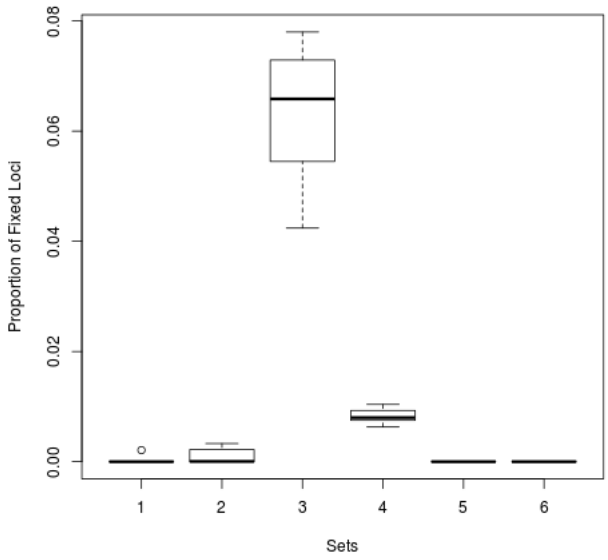


Figure 4.4: Proportion of fixed loci of the sets given in Table 3.1

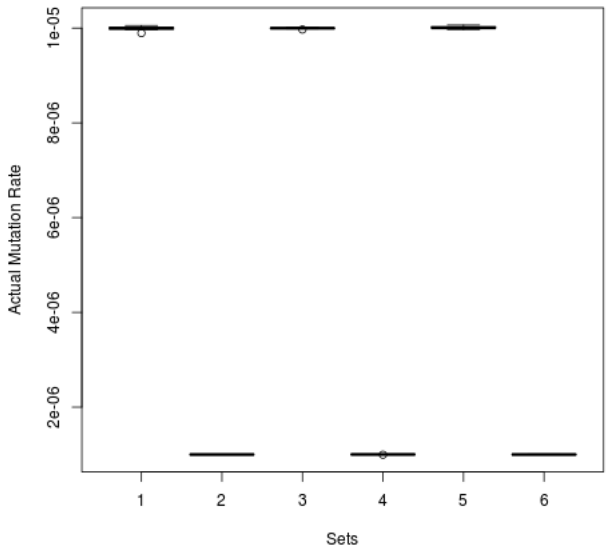


Figure 4.5: Actual mutation rates of the sets given in Table 3.1 that were used in the simulations (see Sec. 3.3.2)

CHAPTER 5

DISCUSSION

In this study, we adopted a mathematical calculation approach, and developed an individual based model to evaluate whether changing parameters as used by Hoggart et al. (2007) could change the genetic makeup of a population. The model is based on a population composed of haploid individuals with their own genetic material. In the simulations, these individuals have a simple life cycle, where they produce offspring which become adults in the next generation. They are simulated in a neutral system, which means that individuals are not under natural or sexual selection. We also assumed that there are no chromosomal mutations, and no recombination. This study is concentrated on two dynamics: the rate of mutation accumulation, and the proportion of polymorphic loci. The former is claimed to be kept unchanged by Hoggart et al. (2007), the latter, we hypothesized as a potentially changing dynamic.

In the calculations we laid out in chapter 2, we showed that the rate of mutation accumulation is only related to μt but not to N, whereas proportion of polymorphic loci is primarily related to $N\mu$, and accumulates with time, but this accumulation is not strictly linear, since the polymorphic loci, increasing by time, decreases the proportion to be added in a single generation. As predicted by the calculations, the results of the simulations were in agreement with Hoggart's rescaling method (2007), the rate of mutation accumulation did not change when parameters N, μ, and t are rescaled; however, the proportion of polymorphic loci have changed (Table 4.1 & Fig. 4.2). It is clear that the Hoggart's rescaling method (2007) changes the genetic makeup of the population, at least, in terms of the proportion of poly-

morphic loci. What is interesting is that the rate of mutation accumulation for sets 1, 2, 4, and 5 were also equal, since their values of μt are equal. So, with or without changing N, one can rescale μ and t, only keeping the rate of mutation accumulation unchanged (comparisons A, B and C in Table 4.1 and Fig 4.2). However, in none of these comparisons (i.e. A, B, and C), the proportion of polymorphic loci could be kept unchanged, which is also an expected result from the equation 2.9, since at least $N\mu$ and/or t values are different between the pairs.

In chapter 2, we hypothesized that the proportion of polymorphic loci can only be kept unchanged, with rescaling N and μ, but by keeping t unchanged. In the table 4.1, the results for the proportion of polymorphic loci for comparisons D and E, show that this hypothesis is also supported. Still in this case, consistent with the predictions, rate of mutation accumulation showed significant difference. In the figure 4.2, it can be seen that Hoggart's rescaling method (2007) (comparison A) decreases the number of polymorphic loci in the rescaled set (i.e. set 1). The relative values of $N\mu$ are equal for these two sets in comparison A, so the expected polymorphism to arise in one generation is the same for both of the simulations, however, the simulation is longer in set 2, therefore more polymorphic loci is expected to accumulate (see Equation 2.9, Table 3.2, and Fig. 4.2).

Similarly, even though, statistically it is concluded that the proportion of polymorphic loci in set 3 is significantly different than that in set 1 ($p= 2.88 \times 10^{-2}$), as well as set 6 ($p= 1.15 \times 10^{-2}$), note that, the p values are slightly below the threshold that, with a 99% confidence interval, we would reject the null hypothesis. There is a strong possibility of type II error in this sense. The point is that set 3 has the highest variance for both the rate of mutation accumulation and the proportion of polymorphic loci (Fig. 4.1). However, it is clear in the figure 4.5 that there is no deviation or high variance in the actual mutation rates used in set 3, so high variance is not due to a mutational bias. This supports the probability that these results could be due to the effect of genetic drift. The effect of drift becomes significant as the number of individuals decreases, due to a "*de facto*" sampling error as Tofanelli et al. (2011) pointed out, and the generation time increases (see Table 3.1). Thus, set 3, having less number of individuals than set 6 and having more number of generations than sets 1 and 6, is subjected to the effect of genetic drift more than these two other sets

and may show a higher variance. This result is consistent with Gavrilets (2005)'s criticism on mutation drift balance.

To understand the effect of drift, a comparison between sets 2, 3, and 4 in terms of the proportion of fixed loci will also be informative (see Fig. 4.4). A pairwise comparison between set 3 and 4, where only μ values are different, it can be seen that increasing mutation rate increases the proportion of fixed loci, which increases the variance of the results as well. If sets 2 and 4, where only N is different, are to be considered, it can be seen that decreased number of individuals increases the probability of fixation, which is consistent with the theory of drift. If the sets 2 and 3, where both N and μ are different, are to be compared, the difference is even more significant. If these conclusions are considered with these results in the figure 4.1, it can be concluded that this high variance is due to disrupted mutation drift balance as pointed out by Gavrilets (2005), since drift acts as a "force" decreasing the variation in the population. The high mutation rate in the set 3 keeps adding more mutations increasing the variation, and obscuring the effect of drift, and resulting in more variation than expected. Here, it is also important that set 3 has also more fixations than other sets, which may be counted as a signal of more drift. Also, note that the main comparison (A), which is Hoggart's rescaling (2007), does not provide enough information on genetic drift. So, it remains as an open question for further work and analysis.

One future direction for this study would be to check how Tajima's D could be affected from a rescaling, in these kind of finite site models. As pointed out by Kim and Wiehe (2009), it is highly dependent on N and it may change due to rescaling. To test this, adding selection to the model would be helpful, since in neutral populations, Tajima's D is expected to fluctuate around zero.

A caveat of this study would be that simulation model developed here is much more simplistic than that of Hoggart et al. (2007)'s. First, the model in this study is haploid, while Hoggart et al. (2007)'s model is diploid and this model lacks recombination and selection, which are present in Hoggart et al. (2007)'s model. However, recombination would affect neither the rate of mutation accumulation nor the proportion of polymorphic loci, since it will neither add a new mutation, nor change

the loci of a present mutation. We also assumed that the loci are neutral. Selection would change the allele frequencies. However, it would only contribute to the already existing bias due to the unequal proportion of polymorphic loci predicted by the neutral model presented here.

Also, the mathematical model did not include the effect of genetic drift. A refinement in the calculations with the effects of genetic drift would solve the problems that might arise from the variation lost due to the effect of drift. Here, one should also note that the generation times of the simulations in this study were not long enough that the populations never come to an equilibrium, and the proportion of fixed loci was very low even for longer simulations. As a possible direction for this study, the equilibrium states can be investigated running longer simulations and refining the mathematical model by adding the effect of genetic drift into these equations. Also to understand the dynamics of mutation drift balance better, the simulations could be monitored across generations instead of recording only the final state.

In conclusion, this study demonstrates that in individual based large-scale genome simulations, the rescaling method proposed by Hoggart et al. (2007) and used by many others (see Subsection 1.1.1) can change proportion of polymorphic loci. The results of this study are consistent with previously mentioned criticisms on this rescaling method (Kim & Wiehe, 2009; Peng & Amos, 2010; Ruths & Nakhleh, 2013; Sargolzaei & Schenkel, 2009; Tofanelli et al., 2011). Adopting algorithmic solutions (as in Ruths and Nakhleh, 2013; Sargolzaei and Schenkel, 2009; and Thornton, 2014) to solve the time and memory problem instead of rescaling method would be a safer approach in large-scale genomic simulations.

REFERENCES

Ayers, K. L., & Cordell, H. J. (2010). SNP selection in genome-wide and candidate gene studies via penalized logistic regression. *Genetic Epidemiology*, *34*(8), 879–91.

Barraclough, T. G., & Vogler, A. P. (2000). Detecting the geographical pattern of speciation from species-level phylogenies. *The American Naturalist*, *155*(4), 419–434.

Bergstrom, C. T., McElhany, P., & Real, L. A. (1999). Transmission bottlenecks as determinants of virulence in rapidly evolving pathogens. *Proceedings of the National Academy of Sciences of the United States of America*, *96*(9), 5095–100.

Birand, A., Vose, A., & Gavrilets, S. (2012). Patterns of species ranges, speciation, and extinction. *The American Naturalist*, *179*(1), 1–21.

Branke, J., Kaussler, T., Smidt, C., & Schmeck, H. (2000). A multi-population approach to dynamic optimization problems. In I. Parmee (Ed.), *Evolutionary design and manufacture* (pp. 299–307). Parmee, I.C..

Brown, D., & Rothery, P. (1993). *Models in biology: mathematics, statistics and computing*. Chichester: John Wiley & Sons.

Carvajal-Rodríguez, A. (2008). Simulation of genomes: A review. *Current Genomics*, *9*(3), 155–159.

Chadeau-Hyam, M., Hoggart, C. J., O'Reilly, P. F., Whittaker, J. C., De Iorio, M., & Balding, D. J. (2008). FREGENE: Simulation of realistic sequence-level data in populations and ascertained samples. *BMC Bioinformatics*, *9*, 364.

Coop, G., Pickrell, J. K., Novembre, J., Kudaravalli, S., Li, J., Absher, D., ... Pritchard, J. K. (2009). The role of geography in human adaptation. *PLoS Genetics*, *5*(6), e1000500.

Cule, E., & De Iorio, M. (2012). A semi-automatic method to guide the choice of

ridge parameter in ridge regression. *arXiv preprint arXiv:1205.0686*, 1–32. doi: 10.1002/gepi.21750

Cule, E., & De Iorio, M. (2013). Ridge regression in prediction problems: automatic choice of the ridge parameter. *Genetic Epidemiology*, *37*(7), 704–14.

Ding, Z., Mailund, T., & Song, Y. S. (2008). Efficient whole-genome association mapping using local phylogenies for unphased genotype data. *Bioinformatics*, *24*(19), 2215–2221.

Duchêne, D. A., Duchêne, S., Holmes, E. C., & Ho, S. Y. W. (2015). Evaluating the adequacy of molecular clock models using posterior predictive simulations. *Molecular Biology and Evolution*, *32*(11), 2986–95.

Duque, T., Samee, M. A. H., Kazemian, M., Pham, H. N., Brodsky, M. H., & Sinha, S. (2014). Simulations of enhancer evolution provide mechanistic insights into gene regulation. *Molecular Biology and Evolution*, *31*(1), 184–200.

Duque, T., & Sinha, S. (2015). What does it take to evolve an enhancer? A simulation-based study of factors influencing the emergence of combinatorial regulation. *Genome Biology and Evolution*, *7*(6), 1415–1431.

Dwyer, G., Levin, S. A., & Buttel, L. (1990). A simulation model of the population dynamics and evolution of myxomatosis. *Ecological Monographs*, *60*(4), 423–447.

Enard, D., Depaulis, F., & Roest Crollius, H. (2010). Human and non-human primate genomes share hotspots of positive selection. *PLoS Genetics*, *6*(2), e1000840.

Fusté, M. C. (2012). *Studies in Population Genetics*. doi: 10.5772/2152

García-Dorado, A., & Gallego, A. (2003). Comparing analysis methods for mutation-accumulation data: a simulation study. *Genetics*, *164*(2), 807–19.

Gavrilets, S. (2005). "Adaptive Speciation"-It is not that easy: Reply to Doebeli et al. *Evolution*, *59*(3), 696–699.

Gavrilets, S. (2014). Models of speciation: Where are we now? *Journal of Heredity*, *105*(S1), 743–755.

Gavrilets, S., Vose, A., Barluenga, M., Salzburger, W., & Meyer, A. (2007). Case studies and mathematical models of ecological speciation. 1. Cichlids in a crater lake. *Molecular Ecology*, *16*(14), 2893–2909.

Gen, M., & Cheng, R. (2000). *Genetic algorithms and engineering optimization*. John Wiley & Sons.

Gillespie, J. H. (1975). Natural selection for within-generation variance in offspring number II. discrete haploid models. *Genetics*, *81*, 403–413.

Griffiths, R. C. (1982). The number of alleles and segregating sites in a sample from the infinite-alleles model. *Advances in Applied Probability*, *14*(2), 225.

Gruijter, D., Maria, J., Lao, O., Vermeulen, M., Xue, Y., Woodwark, C., ... Tyler-Smith, C. (2011). Contrasting signals of positive selection in genes involved in human skin color variation from tests based on SNP scans and resequencing. *Investigative Genetics*, *2*(1), 24.

He, X., Duque, T. S. P. C., & Sinha, S. (2012). Evolutionary origins of transcription factor binding site clusters. *Molecular Biology and Evolution*, *29*(3), 1059–70.

Higgins, K., & Lynch, M. (2001). Metapopulation extinction caused by mutation accumulation. *Proceedings of the National Academy of Sciences of the United States of America*, *98*(5), 2928–2933.

Hitsov, I., Maere, T., De Sitter, K., Dotremont, C., & Nopens, I. (2015). Modelling approaches in membrane distillation: A critical review. *Separation and Purification Technology*, *142*, 48–64.

Hoban, S., Bertorelle, G., & Gaggiotti, O. E. (2012). Computer simulations: Tools for population and evolutionary genetics. *Nature Reviews Genetics*, *13*(2), 110–122.

Hoggart, C. J., Chadeau-Hyam, M., Clark, T. G., Lampariello, R., Whittaker, J. C., De Iorio, M., & Balding, D. J. (2007). Sequence-level population simulations over large genomic regions. *Genetics*, *177*(3), 1725–31.

Hoggart, C. J., Whittaker, J. C., De Iorio, M., & Balding, D. J. (2008). Simultaneous analysis of all SNPs in genome-wide and re-sequencing association studies. *PLoS Genetics*, *4*(7), e1000130.

Howard, R. S., & Lively, C. M. (1994). Parasitism, mutation accumulation and the maintenance of sex. *Nature*, *367*(6463), 554–557.

Hudson, R. R. (1991). Gene genealogies and the coalescent process. In D. J. Futuyma & J. Antonovics (Eds.), *Oxford surveys in evolutionary biology* (Vol. 7, pp. 1–44). New York: Oxford University Press.

Hudson, R. R. (2002). Generating samples under a Wright–Fisher neutral model of genetic variation. *Bioinformatics*, *18*(2), 337–338.

International Human Genome Sequencing Consortium. (2004). Finishing the eu-

chromatic sequence of the human genome. *Nature*, *431*(7011), 931–945.

Jiang, X., Mu, B., Huang, Z., Zhang, M., Wang, X., & Tao, S. (2010). Impacts of mutation effects and population size on mutation rate in asexual populations: A simulation study. *BMC Evolutionary Biology*, *10*, 298.

Jónsson, S. Æ., Staneva, I., Mohanty, S., & Irbäck, A. (2012). Monte Carlo studies of protein aggregation. *Physics Procedia*, *34*(June), 49–54.

Keen, R. E., & Spain, J. D. (1992). *Computer simulation in biology: A BASIC introduction.* (2, illustr ed.; R. E. Keen & J. D. Spain, Eds.). Wiley.

Killcoyne, S., & del Sol, A. (2014). FIGG: Simulating populations of whole genome sequences for heterogeneous data analyses. *BMC Bioinformatics*, *15*, 149.

Kim, Y., & Wiehe, T. (2009). Simulation of DNA sequence evolution under models of recent directional selection. *Briefings in Bioinformatics*, *10*(1), 84–96.

Kimura, M., & Crow, J. F. (1964). The number of alleles that can be maintained in a finite population. *Genetics*, *49*, 725–738.

Kimura, M., & Ohta, T. (1969). The average number of generations until fixation of a mutant mene in a finite population. *Genetics*, *61*, 763–771.

Kimura, M., Ohta, T., & MacArthur, R. H. (1971). *Theoretical aspects of population genetics* (Vol. 4). Princeton: Princeton University Press.

Kirkpatrick, M., & Ravigne, V. (2002). Speciation by natural and sexual selection: Models and experiments. *The American Naturalist*, *159*(S3), S22–S35.

Lanier, H. C., & Knowles, L. L. (2015). Applying species-tree analyses to deep phylogenetic histories: Challenges and potential suggested from a survey of empirical phylogenetic studies. *Molecular Phylogenetics and Evolution*, *83*, 191–9.

Larson-Johnson, K. (2016). Phylogenetic investigation of the complex evolutionary history of dispersal mode and diversification rates across living and fossil Fagales. *The New Phytologist*, *209*(1), 418–35.

Lenski, R. E. (2011). Evolution in action: A 50,000-generation salute to Charles Darwin. *Microbe Magazine*, *6*(1), 30–33.

Lenski, R. E. (2015). *The E. coli long-term experimental evolution project site.* Retrieved from http://myxo.css.msu.edu/ecoli

Lohmueller, K. E. (2014). The impact of population demography and selection on the genetic architecture of complex traits. *PLoS Genetics*, *10*(5), e1004379.

Lohmueller, K. E., Bustamante, C. D., & Clark, A. G. (2011). Detecting directional selection in the presence of recent admixture in African-Americans. *Genetics*, *187*(3), 823–835.

López Herráez, D., Bauchet, M., Tang, K., Theunert, C., Pugach, I., Li, J., ... Stoneking, M. (2009). Genetic variation and recent positive selection in worldwide human populations: Evidence from nearly 1 million SNPs. *PLoS ONE*, *4*(11), e7888.

Lynch, M., Conery, J., & Burger, R. (1995). Mutation accumulation and the extinction of small populations. *The American Naturalist*, *146*(4), 489.

Ma, J., Ratan, A., Raney, B. J., Suh, B. B., Miller, W., & Haussler, D. (2008). The infinite sites model of genome evolution. *Proceedings of the National Academy of Sciences*, *105*(38), 14254–14261.

MacLeod, I. M., Meuwissen, T. H. E., Hayes, B. J., & Goddard, M. E. (2009). A novel predictor of multilocus haplotype homozygosity: Comparison with existing predictors. *Genetics Research*, *91*(06), 413.

Meuwissen, T., & Goddard, M. (2010). Accurate prediction of genetic values for complex traits by whole-genome resequencing. *Genetics*, *185*(2), 623–631.

Morriss-Andrews, A., & Shea, J.-E. (2015). Computational studies of protein aggregation: Methods and applications. *Annual Review of Physical Chemistry*, *66*(1), 643–666.

Muller, H. (1932). Some genetic aspects of sex. *American Naturalist*, *66*(703), 118–138.

Muller, H. (1964). The relation of recombination to mutational advance. *Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis*, *1*(1), 2–9.

Nsoesie, E., Mararthe, M., & Brownstein, J. (2013). Forecasting peaks of seasonal influenza epidemics. *PLoS Currents*, *5*, 1–14.

Oaks, J. R. (2015). Bayesian phylogenetics: Methods, algorithms, and applications. — Edited by Ming-Hui Chen, Lynn Kuo, and Paul O. Lewis. *Systematic Biology*, *64*(6), 1122–1125.

Paquet, E., & Viktor, H. L. (2015). Molecular dynamics, Monte Carlo simulations, and Langevin dynamics: A computational review. *BioMed Research International*, *2015*, 1–18.

Peng, B., & Amos, C. I. (2010). Forward-time simulation of realistic samples for

genome-wide association studies. *BMC Bioinformatics*, *11*, 442.

Peter, B. M., Wegmann, D., & Excoffier, L. (2010). Distinguishing between population bottleneck and population subdivision by a Bayesian model choice procedure. *Molecular Ecology*, *19*(21), 4648–4660.

Pickrell, J. K., Coop, G., Novembre, J., Kudaravalli, S., Li, J. Z., Absher, D., ... Pritchard, J. K. (2009). Signals of recent positive selection in a worldwide sample of human populations. *Genome Research*, *19*(5), 826–37.

Powell, J. E., Visscher, P. M., & Goddard, M. E. (2010). Reconciling the analysis of IBD and IBS in complex trait studies. *Nature Reviews Genetics*, *11*(11), 800–805.

R Core Team. (2015). *R: A language and environment for statistical computing*. Vienna, Austria. Retrieved from https://www.r-project.org/

Ruths, T., & Nakhleh, L. (2013). Boosting forward-time population genetic simulators through genotype compression. *BMC Bioinformatics*, *14*, 192.

Ryman, N., Palm, S., André, C., Carvalho, G. R., Dahlgren, T. G., Jorde, P. E., ... Ruzzante, D. E. (2006). Power for detecting genetic divergence: Differences between statistical methods and marker loci. *Molecular Ecology*, *15*(8), 2031–2045.

Sargolzaei, M., & Schenkel, F. S. (2009). QMSim: A large-scale genome simulator for livestock. *Bioinformatics*, *25*(5), 680–1.

Schaffner, S. F., Foo, C., Gabriel, S., Reich, D., Daly, M. J., & Altshuler, D. (2005). Calibrating a coalescent simulation of human genome sequence variation. *Genome Research*, *15*(11), 1576–83.

Schwartz, R. (2008). *Biological modeling and simulation: A survey of practical models, algorithms, and numerical methods*. London: MIT Press.

Servedio, M. R., Brandvain, Y., Dhole, S., Fitzpatrick, C. L., Goldberg, E. E., Stern, C. A., ... Yeh, D. J. (2014). Not just a theory—The utility of mathematical models in evolutionary biology. *PLoS Biology*, *12*(12), e1002017.

Spencer, C. C. A., Su, Z., Donnelly, P., & Marchini, J. (2009). Designing genome-wide association studies: Sample size, power, imputation, and the choice of genotyping chip. *PLoS Genetics*, *5*(5), e1000477.

Tachmazidou, I., Andrew, T., Verzilli, C. J., Johnson, M. R., & De Iorio, M. (2008). Bayesian survival analysis in genetic association studies. *Bioinformatics*, *24*(18), 2030–2036.

Tachmazidou, I., De Iorio, M., & Dudbridge, F. (2011). Application of the optimal discovery procedure to genetic case-control studies: Comparison with *p* values and asymptotic Bayes factors. *Human Heredity*, *71*(1), 37–49.

Tachmazidou, I., Johnson, M. R., & De Iorio, M. (2010). Bayesian variable selection for survival regression in genetics. *Genetic Epidemiology*, *34*(7), 689–701.

Tajima, F. (1996). Infinite-allele model and infinite-site model in population genetics. *Journal of Genetics*, *75*(1), 27–31.

The International HapMap Consortium. (2005). A haplotype map of the human genome. *Nature*, *437*(7063), 1299–1320.

The International HapMap Consortium. (2007). A second generation human haplotype map of over 3.1 million SNPs. *Nature*, *449*(7164), 851–861.

Thornton, K. R. (2014). A C++ template library for efficient forward-time population genetic simulation of large populations. *Genetics*, *198*(2013), 1–21.

Tofanelli, S., Taglioli, L., Merlitti, D., & Paoli, G. (2011). Tools which simulate the evolution of uni-parentally transmitted elements of the human genome. *Journal of Anthropological Sciences*, *89*, 201–219.

Uricchio, L. H., & Hernandez, R. D. (2014). Robust forward simulations of recurrent hitchhiking. *Genetics*, *197*(1), 221–236.

Varadarajan, A., Bradley, R. K., & Holmes, I. H. (2008). Tools for simulating evolution of aligned genomic regions with integrated parameter estimation. *Genome Biology*, *9*(10), R147.

Vounou, M., Nichols, T. E., & Montana, G. (2010). Discovering genetic associations with high-dimensional neuroimaging phenotypes: A sparse reduced-rank regression approach. *NeuroImage*, *53*(3), 1147–1159.

Watterson, G. (1975). On the number of segregating sites in genetical models without recombination. *Theoretical Population Biology*, *7*(2), 256–276.

White, J. W., Rassweiler, A., Samhouri, J. F., Stier, A. C., & White, C. (2014). Ecologists should not use statistical significance tests to interpret simulation model results. *Oikos*, *123*(4), 385–388.

Winsberg, E. (2015). Computer Simulations in Science. In E. N. Zalta (Ed.), *The stanford encyclopedia of philosophy* (Summer 201 ed.). Retrieved from http://plato.stanford.edu/archives/sum2015/entries/simulations-science/

Wu, C.-C., Shete, S., Chen, W. V., Peng, B., Lee, A. T., Ma, J., ... Amos, C. I.

(2009). Detection of disease-associated deletions in case-control studies using SNP genotypes with application to rheumatoid arthritis. *Human Genetics*, *126*(2), 303–15.

Yan, J., Zhang, H., Du, L., Wernert, E., Saykin, A. J., & Shen, L. (2014). Accelerating sparse canonical correlation analysis for large brain imaging genetics data. *Proceedings of the 2014 Annual Conference on Extreme Science and Engineering Discovery Environment - XSEDE '14*, 1–7.

Yuan, X., Miller, D. J., Zhang, J., Herrington, D., & Wang, Y. (2012). An overview of population genetic data simulation. *Journal of Computational Biology*, *19*(1), 42–54.

Zhang, C. T., & Chou, K. C. (1992). Monte Carlo simulation studies on the prediction of protein folding types from amino acid composition. *Biophysical Journal*, *63*(6), 1523–9.