Stochastic Modeling and Simulation of Biological Phenomena with Applications in Population Genetics and in Cell Populations

by

Biao Li

A THESIS SUBMITTED
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE
Doctor of Philosophy

APPROVED: THESIS COMMITTEE:

Marek Kimmel, Professor, Director
Statistics and Bioengineering, Rice University

Michael W. Deem, Professor, Chair
Bioengineering, Rice University

Dennis D. Cox, Professor
Statistics, Rice University

Oleg A. Igoshin, Associate Professor
Bioengineering, Rice University

Miran Maletic-Savatic, Assistant Professor
Pediatrics, Baylor College of Medicine

Bo Peng, Assistant Professor, Genetics
University of Texas MD Anderson Cancer Center

Houston, Texas
June 2013
Stochastic Modeling and Simulation of Biological Phenomena with Applications in Population Genetics and in Cell Populations

by

Biao Li

Abstract

Stochastic modeling and simulation play important roles in population genetics, statistical genetics, cell biology, molecular biology and evolutionary theory. This thesis explores four aspects of stochastic modeling and simulation of biological phenomena with applications. Research carried out is focused on two major themes. The first one (Chapters 2 and 3) concerns application of stochastic modeling in genetics, specifically to identify biases in analysis of genetics data. The second theme (Chapters 4 and 5) concerns application of Poisson and branching process models to understand various aspects of cell proliferation, using S-phase labeling. In Chapter 2, we investigate factors that are influencing the ascertainment bias of microsatellite allele sizes and explore the impact on estimates of mutation rates. We demonstrate that despite bias, the microsatellite mutation rate estimate in Human exceeds that in Chimpanzee, and also that population bottlenecks and expansions in the recent human history have little impact on the conclusion. Chapter 3 introduces a simulation framework, SimRare, to generate sequence-based data for rare variant association studies and evaluating association test methods. We validate
existing association methods using data generated under various demography and disease models. We demonstrate that the power of each method depends on the underlying model and differences in power between methods are usually modest. Chapter 4 is devoted to the study of early stages of adult hippocampal neurogenesis. We develop a stochastic model to analyze cell labeling data obtained from BrdU labeling experiments. Our model reveals unknown but meaningful biological parameters, such as apoptotic rate and duration time at each stage, etc., to allow us to predict overall efficiency of hippocampal neurogenesis. Chapter 5 focuses on the modeling of DNA replication and bivariate cell labeling experiments. We propose a multiscale modeling of stochastic events related to the measured labeling intensities of both DNA content and replication progression over various exposure times in proliferating cells. We demonstrate that the experimental asymmetry in DNA replication scatterplots is the hallmark of an increasing replication initiation rate in the S-phase of the cell cycle.
Acknowledgments

I would like to express my sincere gratitude to my academic adviser Prof. Marek Kimmel. His profound knowledge, inspiring ideas, stimulating suggestions and continuous support have greatly influenced me through these years. He has been helping me accumulate my research skills and tap into my full potential. Without his wholehearted guidance I might have already lost. I have been truly taking pride in working with him.

I appreciate my Committee members, Prof. Michael W. Deem, Prof. Oleg A. Igoshin, Prof. Dennis D. Cox, Prof. Mirjana Maletic-Savatic and Prof. Bo Peng for the encouragement, inspiration and insightful comments so that I can complete the thesis.

My gratitude goes out to everyone that has helped me in many ways during my research.

Specifically, I would like to thank Prof. Bo Peng and Prof. Christopher I. Amos for providing me with training support in bioinformatics and simulation. I also appreciate very much Dr. Peng for his step-by-step guidance into simuPOP, the fascinating forward-time population genetics simulation environment, which becomes the pivotal tool in my research.

I am grateful to Prof. Mirjana Maletic-Savatic for helping me and getting me familiar with biological and experimental parts in multidisciplinary research topics. I am also indebted to her for the continuous encouragement and support.

I would also like to thank Prof. Suzanne M. Leal for additional mentorship and the opportunity to work collaboratively in her lab to push forward a handful of ideas.
into practice.

I want to thank my friends, schoolmates and colleagues, not only for their help but also the true friendship that walks us together through both difficulties and success. I dedicate this thesis to my family, my wife, my parents and my grandparents. You are the love of my life. Special thanks to my wife, Xu (Wendy) Han, again for her unwavering support and unconditional love in every situation.
## Contents

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Abstract</td>
<td>i</td>
</tr>
<tr>
<td>Acknowledgments</td>
<td>iii</td>
</tr>
<tr>
<td>List of Figures</td>
<td>xiii</td>
</tr>
<tr>
<td>List of Tables</td>
<td>xiv</td>
</tr>
<tr>
<td><strong>1 Introduction</strong></td>
<td>1</td>
</tr>
<tr>
<td>1.1 The role of stochastic modeling and simulation</td>
<td>7</td>
</tr>
<tr>
<td>1.1.1 Deterministic model vs stochastic model</td>
<td>8</td>
</tr>
<tr>
<td>1.1.2 Characteristics of stochastic models</td>
<td>8</td>
</tr>
<tr>
<td>1.2 Background in stochastic processes</td>
<td>9</td>
</tr>
<tr>
<td>1.2.1 Probability generating function</td>
<td>10</td>
</tr>
<tr>
<td>1.2.2 Bernoulli process</td>
<td>10</td>
</tr>
<tr>
<td>1.2.3 Counting process</td>
<td>10</td>
</tr>
<tr>
<td>1.2.4 Poisson process</td>
<td>11</td>
</tr>
<tr>
<td>1.2.4.1 Compound Poisson process</td>
<td>12</td>
</tr>
<tr>
<td>1.2.4.2 Non-homogeneous Poisson process</td>
<td>12</td>
</tr>
<tr>
<td>1.2.5 Markov chain</td>
<td>13</td>
</tr>
<tr>
<td>1.2.6 Branching process</td>
<td>13</td>
</tr>
<tr>
<td>1.2.6.1 Bellman-Harris (multiple-type) branching process</td>
<td>14</td>
</tr>
</tbody>
</table>
2 Ascertainment bias in discovery of microsatellite loci

2.1 Motivation and overview

2.2 Introduction

2.3 Modeling and simulation

2.3.1 Evolution of a DNA-repeat locus

2.3.2 Conditional distributions and ascertainment bias of allele sizes

2.3.2.1 Chromosomes sampled from populations 1 and 2

2.3.2.2 Chromosomes sampled from population 1
2.3.2.3 Probability generating functions and expectations of incremental
changes of allele sizes ........................................ 35
2.3.3 Mean allele size difference, $D$ ................................ 36
2.3.4 Derivation of the analytical solution to $D$ .................. 38
  2.3.4.1 Derivation of equation (2.4) ............................ 38
  2.3.4.2 Derivation of equations (2.5) and (2.6) ................. 39
  2.3.4.3 Derivation of equations (2.7) and (2.8) ................. 39
  2.3.4.4 Derivation of computational expressions for equations (2.10) and (2.11) ........................................ 40
  2.3.4.5 Derivation of computational expression for equation (2.12) ........................................ 44
  2.3.4.6 Derivation of equations (2.13) and (2.14) ............... 44
2.3.5 Simulation method ............................................. 45
2.4 Results ............................................................. 46
  2.4.1 Summary of modeling results ............................... 46
  2.4.2 Genetic algorithm and plausible parameter ranges ............ 49
  2.4.3 Comparisons of empirical statistics derived from human and chimpanzee microsatellite data .......................... 51
    2.4.3.1 Remark - derivation of $P(L \geq x)$ ...................... 52
  2.4.4 Influence of bottlenecks and expansions in human history ............... 54
    2.4.4.1 Remark - derivation of mean allele size difference, $D$, in the extended model .................................... 57
2.5 Discussion and conclusions .................................... 58
3 Simulation framework to generate and analyze sequence-based data for rare variant association studies .......... 64
  3.1 Motivation and overview ...................................... 64
  3.2 Introduction .................................................. 65
3.3 Description .................................................. 66
  3.3.1 Generation of Variant Data ............................... 66
  3.3.2 Generation of Phenotype Data ......................... 67
  3.3.3 Analysis of Data and Evaluation of Type I and II error 68
  3.3.4 Program overview ..................................... 68
    3.3.4.1 Where to apply ................................. 69

3.4 Program Guide ........................................... 70
  3.4.1 Module of variant data generation ......................... 70
    3.4.1.1 Basics ......................................... 70
    3.4.1.2 Demography .................................. 72
    3.4.1.3 Genetic forces .............................. 73
    3.4.1.4 Save and run ................................ 75
    3.4.1.5 Output files ................................ 76
  3.4.2 Module of genotype and phenotype generation ............. 77
    3.4.2.1 Initialization ................................ 77
    3.4.2.2 Model parameters ............................ 79
    3.4.2.3 Mimic genotyping ............................ 83
    3.4.2.4 Save and run ................................ 83
  3.4.3 Module of association tests ............................. 84
    3.4.3.1 Configuration ................................ 84
    3.4.3.2 Existing methods ............................ 85
    3.4.3.3 Novel method ................................ 86
    3.4.3.4 Output figures and files ....................... 87
  3.5 Case-control sample rare variant association analysis - an example to use SimRare ................. 88
    3.5.1 Create variant data pool ............................ 88
    3.5.2 Specify disease model parameters under $H_1$ .................. 89
3.5.3 Specify disease model parameters under $H_0$ ............................... 91
3.5.4 Perform association tests ......................................................... 91
3.5.5 Results .................................................................................. 92

3.6 Power comparison of existing complex trait rare variant association
methods ..................................................................................... 94

3.6.1 Simulation setting ................................................................. 95
3.6.2 Case-control prevalence model .............................................. 96
3.6.3 Case-control population attributable risk model ...................... 96
3.6.4 Results .................................................................................. 99

4 Modeling of adult hippocampal neurogenesis in its early stages 101

4.1 Motivation and overview .......................................................... 101
4.2 Introduction .............................................................................. 102
4.3 Materials and methods ............................................................. 105
4.3.1 Animals .............................................................................. 105
4.3.2 Histology ............................................................................ 106
4.3.3 Confocal Microscopy ............................................................. 107
4.3.4 Cell transit times ................................................................. 107
4.3.5 Classification of cell stages ................................................... 108
4.3.6 Expected number of ANP divisions ....................................... 111
4.3.7 Modeling of the neurogenic cascade using branching process ... 111
4.3.8 Modeling and simulation of cell labeling curves ...................... 116
4.3.9 Parameter search and goodness-of-fit ..................................... 117
4.4 Results and discussion ............................................................ 118
4.4.1 Data .................................................................................... 118
4.4.2 Data cleaning and transformation .......................................... 120
4.4.2.1 Adjust cell proportions ..................................................... 120
5  Modeling of stochastic events related to DNA content distributions in proliferating cells

5.1 Motivation and overview

5.2 Introduction

5.3 Methods

5.3.1 Laboratory procedures

5.3.2 Computations

5.3.2.1 Transformation of LSC data

5.3.2.2 Modeling and simulation steps

5.3.2.3 Cell transit times

5.3.2.4 Modeling and simulation of in vitro cell culture

5.3.2.5 Modeling DNA replication

5.3.2.6 Modeling and simulation of cell labeling experiments

5.3.3 Simulation program

5.4 Results

5.4.1 Summary of LSC observations

5.4.2 Exploratory simulations

5.4.3 Simulations based on independently estimated intensity rate of the DNA synthesis NHPP

5.4.4 Analysis of DNA content frequency histogram
# List of Figures

2.1 Evolutionary history of a locus in two species ........................................ 31

2.2 Flowchart illustrating key steps of simulation in the study of ascertainment bias using simuPOP ................................................................. 45

2.3 Observed difference $D$ in allele sizes may be positive or negative ............ 48

2.4 Schematic representation of human demographic history with recent bottlenecks and expansions ................................................................. 55

3.1 Snapshot of SimRare interface - module of generating variants data pool ...... 71

3.2 Snapshot of SimRare interface - module of generating genotype & phenotype data ................................................................. 78

3.3 Snapshot of SimRare interface - module of association tests ....................... 85

3.4 Snapshot of SimRare interface - create variant data pool .......................... 89

3.5 Specify case-control disease model parameters under the alternative hypothesis ($H_1$) ................................................................. 90

3.6 Evaluate rare variant association test methods .......................................... 93

3.7 Power comparison of association tests under both $H_0$ and $H_1$ ............... 93

3.8 Quantile-Quantile plots of CMC-one method under both $H_0$ and $H_1$ ....... 94

3.9 Power comparison with impact of non-causal variants ............................... 97

3.10 Power comparison with detrimental and protective variants ....................... 98

3.11 Power comparison with attributable risks ............................................ 99

4.1 Schematic representation of current knowledge about hippocampal neurogenic cascade ................................................................. 104
4.2 Paradigm of BrdU (Bromodeoxyuridine) labeling experiments

4.3 Branching process modeling of the neurogenic cascade

4.4 An illustration of simulating the dynamic of a neural progenitor cell

4.5 Fitting model to experimental curves - data vs simulation results

4.6 Residuals plot - data vs simulation results

4.7 Estimated apoptosis rate at each cell stage that yields best fit to data

4.8 Prediction of dynamics of neurogenesis under reduced apoptosis

5.1 Bivariate DNA content versus EdU incorporation distributions

5.2 Graphical user interface of simulation program for DNA replication and labeling experiments

5.3 Illustration of the linear functions as rate function of NHPP

5.4 Simulation results with constant rate function

5.5 Comparison of simulation results taking linear decreasing rate functions

5.6 Comparison of simulation results assuming linear increasing rate functions

5.7 Replication activity profile derived from replication timing map

5.8 Simulation results via replication rate function derived from the replication activity profile

5.9 Comparison of DNA content distributions between data and simulation results

5.10 Characteristic pattern of replication already shown for short EdU incubation period
List of Tables

2.1 Parameter settings that yield a good fit, for a range of realistic effective population sizes and mutation rates ........................................ 50

4.1 Model parameters that can be estimated experimentally or not ........................................ 118

4.2 Data of total BrdU+ cell count and BrdU+ apoptotic cell count ........................................ 119

4.3 Data of estimated proportion of BrdU+ cell of each type ........................................ 120

4.4 Cleaned data of estimated proportion of BrdU+ cell of each type ........................................ 122

4.5 Summary of estimated cell counts in each type ........................................ 124

4.6 List of model parameters and ranges of possible values ........................................ 127

4.7 Parameter estimates that yield best fit and comparison with estimates in literature ............ 131

5.1 Input parameters for simulating in vitro cell culture ........................................ 150

5.2 Input parameters for simulating labeling experiments of DNA contents .......................... 150
Chapter 1

Introduction

There is an increasing awareness of the important role of stochasticity playing in biological systems. Many processes in population genetics and cell population dynamics are comprised of random components. In this thesis, we investigate biological phenomena in four aspects using stochastic modeling and simulation with applications. This thesis results in stochastic models and simulation programs that will benefit future research in respective areas, particularly in those that are concerned about sample selection bias in the analysis of genetic data and cell population dynamics using S-phase labeling. Applying these stochastic models and simulation programs in real data we also investigate unknown but meaningful biological parameters, that will lead to future development of novel experimental designs and deeper understanding of the underlying biological systems.

In the first study (chapter 2), we design and develop a stochastic model to characterize the ascertainment bias in the interspecies comparison on microsatellite loci. Such bias was first discovered in the comparison of microsatellite lengths between human and chimpanzees when the loci were originally discovered in human genome.
The same observation of ascertainment bias also occurs while we compare homologous single nucleotide polymorphisms (SNPs) loci in different human populations or different related species. Such phenomenon is rooted in the fact of us tending to choose the locus of interest in a biased manner. Although previous studies agreed that the ascertainment bias is common to observe, its explicit mechanisms is still obscure. Reciprocal studies and genome-wide based research show that such variations are due to the combined effect of biased sampling and intrinsic systematic difference, e.g. evolutionary forces, such as genetic drift, demography and mutation rates. Based on such finding, we build a more realistic mathematical model which can both explore complex evolutionary history and clearly discern the sampling effect in inter-species studies. Using the model to re-analyze literature data on allele size difference on microsatellite loci between Human and Chimpanzee, we discover that despite ascertainment bias, mutation rate estimate in Human exceeds or equals to that in Chimpanzee. We also demonstrate that population bottlenecks and expansions in recent human history have little impact on such estimates. Practically, our modeling framework can be easily adapted or extended by further research to explore any arbitrarily complex evolutionary scenarios. We also implement a simulation program using simuPOP to simulate evolutionary trajectories rapidly and accurately in case that derivation to analytical solution becomes too complex to attain. This research will benefit geneticists to fairly estimate genetic and evolutionary parameters when ascertainment bias is present during sample selection procedure.

In the second part (chapter 3), we build a stochastic simulation program, SimRare,
which integrates generation of rare variant genotype/phenotype data and evaluation of association methods using a unified platform. Currently, it is difficult to evaluate complex trait rare variant association methods because there is no standard to generate data and often comparisons are biased in order to make a particular method to appear superior to others. Therefore, we design and develop SimRare, which can generate data using realistic population demographic and phenotypic models, and evaluate association methods, including novel method, in an unbiased manner. Using SimRare we validate existing association methods with data generated under various scenarios of demographic history and disease etiology. We demonstrate that the power of each method depends on the underlying model and differences in power between methods are usually modest. SimRare, as a stochastic simulation platform with integrated modules of data generation and statistical analysis, can be widely used at many different levels for rare variant association studies. Practically SimRare is distinguished as benchmark software to impartially assess the power and robustness of novel and existing rare variant association methods under a broad range of contexts. This feature should greatly aid researchers in the development of methods to test for rare variant associations. SimRare can also assist genetic epidemiological study designs by allowing a wide variety of scenarios to be tested. For instance, for a population-based case-control exome-sequencing study SimRare can be used to determine the necessary sample size to achieve adequate power under a variety of different phenotype models. SimRare will greatly increase the ability and speed at which researchers can develop association methods. SimRare allows researchers to create association methods without having to develop
special software to generate data in order to evaluate their methods. It also fa-
cilitates an easy way to compare their method to previously developed methods,
since evaluation can be performed on consistent data sets and it is not necessary to
develop software to implement a large variety of methods for comparison purposes.
Ultimately, this universally functional software equipped with user-friendly inter-
face will aid genetic epidemiologists and statistical genetics in the development of
methods and the design of sequence based genetic association studies.

The third part of this thesis (chapter 4) focuses on the the modeling of adult hipp-
ocampal neurogenesis in its early stages. Numerous studies have indicated that
neurogenesis is important for learning and memory. The principle purpose of adult
neurogenesis is to generate new neurons required for establishment of new memo-
ries, particularly related to spatial learning. In the adult mammalian brain, new
neurons are produced from neural progenitor cells (NPC) which reside in the sub-
ventricular zone (SVZ) and in the subgranular zone (SGZ) of the dentate gyrus
(DG). In the SGZ, the NPC include both quiescent and amplifying neuroprogeni-
tors (QNP and ANPs). QNP, named because of their low rate of division, divide
asymmetrically and give rise to ANPs. The ANPs proliferate actively, amplifying
their number by symmetric divisions. Eventually, they divide asymmetrically and
transition into postmitotic neuroblasts (NBs) which slowly migrate into the granule
cell layer. There, NB differentiate into immature neurons and develop functional
synaptic connections, finally becoming fully integrated, mature granule cells. Im-
nunohistochemically, we can precisely and reliably quantify each cell type within
the neurogenic cascade. Early stages of neurogenesis are very complex, since mech-
anisms that determine cell proliferation, differentiation, and death are all in effect at the same time. As experimental studies of such a complex system require years of work, we develop a computational model that will provide us with a unique tool to quickly and reliably test a variety of scenarios affecting ANP and NB fates within the neurogenic cascade, to predict the outcomes of minute changes of each step of the neurogenesis. In particular, there were no computational models that take into account apoptosis as one of the possible fates of cellular components of neurogenic cascade. Therefore, we design a comprehensive computational model of early neurogenesis, based on our experimental data, utilizing branching stochastic processes and simulation. Based on current data of single BrdU pulse-and-chase labeling curves from animal brains (1 month old mice), our modeling and simulation results reveal unknown but meaningful biological parameters, among which the most crucial ones are apoptotic rates at different cell stages. This study concludes that apoptotic rates reach maximum for NBs, renewal probability of ANPs is low, variance of cell durations of NBs is high and expected duration for apoptotic cells is short, etc. Practically, the established stochastic modeling and simulation framework will enable us to manipulate neurogenesis and enhance the number of viable neurons. It can also generate predictions of the behavior of the neurogenic system under perturbations such as increase or decrease of apoptosis due to disease or treatment, respectively. Eventually, this research will establish anatomical, functional, and computational foundations for our future studies aimed to manipulate newborn cell apoptosis and phagocytosis. The ability to increase the number of viable neurons may have significant therapeutic implications for drug addiction, learning
impairments, and mood disorders as all these conditions have been associated with decreased neurogenesis.

In the fourth part (chapter 5), we demonstrate that experimental asymmetry in DNA-replication scatterplots is the hallmark of an increasing replication initiation rate in the S-phase of the cell cycle. Living cell needs to produce two identical copies of its DNA sequences before splitting into two daughter cells; the process is called DNA replication. In eukaryotic cells, DNA replication begins at multiple points of the genome, the so-called replication origins. Across entire human genome, there are tens of thousands of such origins at specific locations. However, they are fired at different times and at present little is known about the mechanisms coordinating the timing program. It is suggested that the normal replication timing might be altered or disrupted when a healthy cell becomes cancerous or diseased. Therefore, revealing how firing times of origins regulate DNA replication process may be useful both in understanding the temporal order of replication and in understanding abnormal replication in diseased cells. Stochastic dynamics of molecular level events in cells such as DNA replication is still not directly observable. As DNA synthesis is one of the key phases of cell cycle progression, understanding its kinetics will allow gaining insights into mechanisms of normal and cancer cells growth. We build a stochastic modeling and simulation framework which provides a tool for connecting the molecular events with single-cell characteristics observable by assays based on laser scanning cytometry (LSC). In the present study we labeled newly synthesized DNA in A549 human lung carcinoma cells with 15 - 120 min pulses of EdU. All DNA was stained with DAPI and cellular fluorescence was measured by LSC. Our
current finding proves consistency between the S-phase DNA-replication rate based on molecular-scale analyses, and cell population kinetics ascertained from EdU/DAPI scatterplots. Practically, our stochastic model and simulation program can be used by both specialists and non-specialists to pursue a quantitative understanding of the relation between different cell proliferation parameters and the observed cell labeling intensities. Our approach also opens a possibility of similar modeling to study the effect of anticancer drugs on DNA replication/cell cycle progression and also to quantify other kinetic events that can be measured during S-phase.

In this chapter, I will first discuss the role and characteristics of stochastic modeling and simulation, comparison of deterministic and stochastic models, and the advantages of using stochastic modeling in the situation that deterministic modeling technique is unsuitable. Secondly, I will provide the basic background knowledge covering a variety of subjects that are required as prerequisites to read and understand this dissertation.

1.1 The role of stochastic modeling and simulation

Biological systems are complex and involving a great number of entities that interact and determine the dynamics of the objects taken as consideration. Computational techniques use modeling and simulation approaches to represent, decompose and understand such complexity in biologically meaningful but simpler components. A deterministic model is often used if the result is determined by a series of conditions in a strict and straightforward manner, whereas a stochastic model is applied if the cause and effect are randomly or stochastically related.
1.1.1 Deterministic model vs stochastic model

The deterministic modeling process often uses a set of ordinary differential equations (ODEs) or partial differential equations (PDEs). The equation set is typically solvable analytically or numerically. Differential equation-based approach has been widely used to model and simulate biological systems. Although these techniques are extremely powerful, they have limitations such as,

- when a tractable solution cannot be obtained from the set of equations (Chaturvedi et al., 2005) [20];
- when there is lack of available data and information to formulate a numerically deterministic model;
- when the model assumption does not hold if the system is actually evolving in pure or semi-random manner (Dhar et al., 2004) [34].

Stochastic modeling is distinguished from the equation-based deterministic modeling. A stochastic model is generally not solved analytically. In stochastic simulation random numbers are usually generated. Such simulation scheme using the random number generator is named Monte Carlo method or Monte Carlo simulation. Arguably, the stochastic model is more informative than the deterministic model since the stochastic one captures heterogeneity due to the random nature of the biological system under study.

1.1.2 Characteristics of stochastic models

The following features are characteristic of stochastic models:
• model contains random components regarding the spatial-temporal order and behavior of biological phenomena;

• a single simulation run yields only one possible result, but repeated simulation with the same parameter values can lead to different outputs;

• multiple runs that generate a large number of replicates are required to estimate probability distributions and expected average outcomes.

Stochastic modeling and simulation are commonly used to (Chen et al., 2006) [22]:

• clarify and make explicit the different hypotheses or assumptions underlying a conceptual biological model;

• aid biologists to formalize biological evidence into the computational model;

• describe the modeling system by various types of stochastic processes, such as renewal process, Poisson process, Markov process, Branching process, etc.;

• implement the experimental procedure in silico and generate the experimental results by simulation;

• estimate unknown but meaningful biological parameters;

• validate the biological model whose dynamics can not be directly observed, check whether data from simulation are consistent with biological data;

• make predictions from simulation data and assist in design of experiments.

1.2 Background in stochastic processes

A stochastic process $X = \{X(t), t \in T\}$ is a collection of random variables, where
$T$ is an index set of time $t$ and $X(t)$ is a random variable denoting the state of the process at time $t$. $X$ is called a discrete-time stochastic process or continuous-time process if the index set $T$ is a countable set or a continuum, respectively (Ross, 1996) [101].

1.2.1 Probability generating function

Firstly, we introduce the probability generating function (pgf) because it is an important tool in studies using discrete-time stochastic processes. The pgf is defined as the representation of power series of the probability mass function (pmf) of a discrete random variable. If we denote $X$ as a discrete random variable which takes values on the set of non-negative integers and $\{p_k := P(X = k), k = 0, 1, \cdots \}$ as its pmf, then we have the pgf of $X$ defined as

$$f(s) = E[s^X] = \sum_{k=0}^{\infty} p_k s^k, |s| \leq 1$$

Some simple properties of pgfs are $f(1^-) = 1$, $E[X] = f'(1^-)$ and $p_k = \frac{f^{(k)}(0)}{k!}$.

1.2.2 Bernoulli process

A Bernoulli process is a discrete-time stochastic process which consists of a sequence of independent random variables that are valued by one out of two symbols. Practically, a consecutive trails of coin flipping mimics a Bernoulli process. A random variable in such a sequence is called a Bernoulli variable.

1.2.3 Counting process

A stochastic process $\{N(t), t \geq 0\}$ is called a counting process if $N(t)$ represents
the total number of events that have occurred up to time \( t \). A counting process
must satisfy: (i) \( N(t) \geq 0 \); (ii) \( N(t) \) is integer valued; (iii) if \( s < t \), \( N(s) \leq N(t) \);
(iv) for \( s < t \), \( N(t) - N(s) \) equals the number of events that have occurred in the
interval \([s, t]\).

A counting process is said to possess independent increments if the numbers of
events that occur in disjoint time intervals are independent, e.g. \( N(s) \perp \perp N(t) - N(s) \), where \( 0 < s < t \). Also, a counting process is said to possess stationary
increments if the distribution of the number of events that occur in any interval of
time depends only on the length of the time interval, e.g. \( N(t_2 + s) - N(t_1 + s) \) has
the same distribution as \( N(t_2) - N(t_1) \) does, where \( 0 < t_1 < t_2 \) and \( 0 < s \).

### 1.2.4 Poisson process

One of the most important types of counting processes is Poisson process. If we
define the counting process \( \{N(t), t \geq 0\} \) with following properties: (i) \( N(0) = 0 \); (ii) the process has independent increments; (iii) the number of events in any
interval of length \( t \) is a Poisson distribution with mean \( \lambda t \), (that is, for all \( s, t \geq 0 \),
\( P(N(t + s) - N(s) = n) = e^{-\lambda t} \frac{(\lambda t)^n}{n!}, n = 0, 1, 2, ... \)), such a counting process is said
to be a Poisson process having rate \( \lambda \) (\( \lambda > 0 \)).

Since the distribution is independent of \( s \), it follows that the Poisson process has
stationary increments, that is \( P(N(t + s) - N(s) = n) = P(N(t) = n) \).
1.2.4.1 Compound Poisson process

A compound Poisson process with rate $\lambda > 0$ and jump size distribution $X$ is a continuous-time stochastic process $\{Y(t) : t \geq 0\}$ given by

$$Y(t) = \sum_{i=1}^{N(t)} X_i$$

where, $\{N(t) : t \geq 0\}$ is a Poisson process with rate $\lambda$, and $\{X_i : i \geq 1\}$ are independent and identically distributed (iid) random variables, with distribution function $X$, which are also independent of $\{N(t) : t \geq 0\}$.

Useful properties are $E[Y(t)] = E[E[Y(t)|N(t)]] = E[N(t)]E[X] = \lambda t E[X]$, $f_{Y(t)}(s) = f_{N(t)}(f_X(s)) = e^{\lambda t(f_X(s)-1)}$, where $f_{Y(t)}$ is the pgf of $Y(t)$, $f_{N(t)}$ is the pgf of $N(t)$ and $f_X$ is the pdf of $X$.

1.2.4.2 Non-homogeneous Poisson process

A counting process $\{N(t), t \geq 0\}$ is said to be a nonhomogeneous Poisson process with intensity function $\lambda(t), t \geq 0$ if: (i) $N(0) = 0$; (ii) $\{N(t), t \geq 0\}$ has independent increments; (iii) $P(N(t+h) - N(t) \geq 2) = o(h)$; (iv) $P(N(t+h) - N(t) = 1) = \lambda(t)h + o(h)$. If we let

$$m(t) = \int_0^t \lambda(s)ds$$

it can be shown that $N(t+s) - N(t)$ is a Poisson distributed random variable with expected value $m(t+s) - m(t)$. 
1.2.5 Markov chain

For a stochastic process \( \{X_n, n = 0, 1, 2, \ldots\} \) that takes on a finite or countable number of possible values, if the conditional distribution of any future state \( X_{n+1} \), given the past states \( X_0, X_1, \ldots, X_{n-1} \) and the present state \( X_n \), is independent of all past states and only dependent on the present state, such process is called a Markov chain. Hence, if the process is said to be in state \( i \) at time \( n \), denoted by \( X_n = i \), then the probability that the process will be in state \( j \) in the next step depends only on \( X_n = i \), that is,

\[
P( X_{n+1} = j | X_n = i, X_{n-1} = i_{n-1}, ..., X_1 = i_1, X_0 = i_0 ) = P( X_{n+1} = j | X_n = i )
\]

1.2.6 Branching process

Branching process theory can be traced to the social scientists in the 19th century studying the extinction of family lines. From that time on, a large number of biological problems have been modeled by branching processes, particularly in the analysis of evolutionary cell population and population genetics. For example, during the evolution of a population of some reproducing particles, each particle lives for a life time, independently of the others, and produces a random number of new offspring. If each particle lives for a fixed time unit and only upon death it produces progeny, then the process is called a Galton-Watson branching process. If each particle has an exponentially distributed lifetime independent of the offspring distribution, then the process is called a Markov continuous-time branching process. If the life time of each particle is subject to a random variable with an arbitrary
distribution, independent of life times of the offspring, then this process is named as an age-dependent (Bellman-Harris) branching process (Kimmel and Axelrod, 2002).

1.2.6.1 Bellman-Harris (multiple-type) branching process

Bellman-Harris process assumes that particles live independently for random periods of time and produce random numbers of new particles at the end of their life time. \( G(t) \) denotes the distribution functions of life times of individuals particles and \( h(s) \) represents the generating function of the number of progeny of one particle. If the age of the particle is 0 at time \( t = 0 \), the probability generating function of \( \mu(t) \), the number of particles, satisfies the equation that:

\[
F(s, t) = \int_0^t h[F(s, t - u)] dG(u) + s [1 - G(t)]
\]

1.3 Background in population genetics

Population genetics was primarily focused on the study of genetic variation between and within populations, e.g. species, organisms, etc., over generations and over time. It was mainly a theoretical discipline and only until recently begins its data and computation driven trend due to the rapid development of informatic technology and simulation programs (Neale et al., 2008). Population genetics also includes the study of various forces that result in evolutionary changes in species through time. Each individual in the species is characterized by its genotype - the genetic constitution, and phenotype - the manifested trait. Under most biological situations, the relationship is complex between genotype and phenotype
1.3.1 Population demography

Population demography pertains to the study of size, structure and distribution of human populations. Its analysis includes the effects of both genetic and environmental factors on the population dynamics over time and (or) space. In population genetics, demography defines the underlying population structure, and temporal and spatial fluctuation in the population size.

For the purpose of theoretical modeling and (or) simulation, we often use effective population size to denote the population demography. An effective population size is defined as the number of individuals in an idealized population that has a value of any given population genetic quantity that is equal to the value of that quantity in the population of interest.

1.3.2 Genetic forces

The major forces that cause genetic variation in populations are:

- genetic drift - the random variation in allele frequency due to sampling effect in small populations;
- mating scheme - rules that manipulate how individuals are mated in the parental generation to reproduce the offspring generation;
- mutation - any change of the nucleotide sequence of the genome;
- natural selection - a natural process that favors or induces survival and perpetuation of one kind of organism over the others;
• Recombination - the natural formation of offspring of genetic combinations not present in parents, by the processes of crossing over or independent assortment.

1.3.3 Wright-Fisher model

Wright-Fisher model describes the process of genetic drift in a finite population (Wright, 1931; Fisher, 1990). Model’s assumptions are: $N$ diploid individuals, non-overlapping generations, monoecious reproduction with an infinite number of gametes, random mating, no mutation or selection. Each parent has an equal probability of contributing a gamete to a progeny that will live in the next generation. Given two allele $A_1$ and $A_2$ in the population, each generation will experience genetic drift through random sampling. Due to the assumption of infinite number of gametes, the transition probability from $i$ copies of $A_1$ to $j$ copies of $A_1$ in the next generation is given by the binomial probability distribution

$$P_{ij} = \binom{2N}{i} \left( \frac{i}{2N} \right)^j \left( 1 - \frac{i}{2N} \right)^{2N-j}$$

If $p$ is the frequency of $A_1$ and $q$ is the frequency of $A_2$, the expected frequency of individuals with genotype $A_1A_1$, $A_1A_2$, $A_2A_2$ in the next generation will be $p^2$, $2pq$, $q^2$.

1.3.4 The coalescence

The basic coalescent theory describes the genealogical relationships among individuals in a Wright-Fisher population (Kingman, 1982; Kingman, 2000). Given a sample of individuals at present, the coalescent model starts from finding the common ancestor of two closest related individuals, then traces all individuals
including those being found as common ancestors in the backward sense of time and ends in a single ancestor, the most recent common ancestor, which all individuals relate to.

If we record step by step connecting any two related individuals to their common ancestors with lines and marking themselves with nodes, the process of tracing the most recent common ancestor forms a phylogenetic tree. For instance, if only two individuals are sampled at present, they have a common ancestor $T$ units of time before present, where $T$ has an exponential distribution with parameter $1/2N$, $N$ is the effective population size. In the case of $n$ alleles, the expected value of the common ancestor time is given by $E[T_n] = \frac{4N}{n(n-1)}$.

The effect of genetic drift can also be well modeled by the coalescence with identical conclusion drawn as that in the forward-time manner (Gillespie, 2004). In simple cases, the classic coalescence assumes no other genetic factors or evolutionary interests, such as recombination, mutation, gene flow, demographic history, selection and so on. However, the development of advanced mathematical modeling in coalescent theory allows arbitrarily complex evolutionary scenarios to be examined analytically (Leach et al., 2009; Siepel, 2009).

### 1.3.5 Common and rare variants

A single nucleotide polymorphism (SNP) is a DNA sequence variation occurring at a single nucleotide site, when A, T, C, or G differs between paired chromosomes of individuals in the population.

Common variants are SNP sites with minor allele frequencies (MAFs) greater than 1% (or 5%), and rare variants are those with MAFs smaller than or equal to 1%
1.3.6 Simple and complex genetic diseases

Disease alleles are inherited from parents to offspring, but are not always leading to the onset of phenotypic trait (disease). If a disease can be caused by a single inherited defective gene (or disease locus), then such disease is called simple disease. Simple diseases are monogenic and usually have obvious family inheritance pattern. If a disease is due to a combination of multiple genes and environmental factors, and (or) interactions between them, then such disease is a complex or multifactorial disease. Most commonly known human diseases are complex, such as heart disease, high blood pressure, diabetes, cancer, obesity and Alzheimer’s.
1.4 Background in cell populations and labeling techniques

Understanding of cell population dynamics becomes impossible without quantitative estimates of cellular activities. Labeling techniques become the most widely used approaches to measure the single-cell characteristics and collectively reflect the systematic behavior on the cell population level.

1.4.1 Cell proliferation, differentiation and apoptosis

Cell proliferation denotes cell growth and cell division or reproduction. When a cell proliferates it grows, divides and gives rise to two daughter cells. If a cell population is composed of proliferating cells only, each generation of cells should be twice as numerous as the previous generation.

Cell differentiation is the normal process by which a less specialized cell develops or matures to possess a more specialized and distinct form and function. For example, a single-celled zygote can develop into a multicellular embryo which will be able to further develop and differentiate into a more complex system of distinct types of cells as a fetus. Most of mammalian differentiating cells are derived from stem cells and unidirectional. Once a cell differentiates it stays differentiated and can not be converted to its original form.

Apoptosis, or programmed cell death is the process that the cell signals itself to commit suicide. There are multiple situations under which cells may need to be destroyed, e.g. damaging, infection, malfunction and programmed termination.

1.4.2 Cell cycle stages and transit times

The cell division cycle defines a series of events that take place to lead a proliferating
cell to replicate its genome and divide into two progeny. The entire process undergoes three distinct stages (or phases), including $G_1$ - a preparation stage when cell size increases, $S$ - a synthesis stage when DNA replication occurs, $G_2M$ - a combined stage of $G_2$-phase and $M$-phase when the cell ensures that it is ready to divide and then completes the division.

The amount of time that a cell spends in a stage $a$ before entering the next stage $b$ is defined as the cell transit time from stage $a$ to $b$ or the cell duration at $a$.

1.4.3 Cell labeling techniques

Using fluorescent markers in confocal microscopy and laser scanning cytometry, cell labeling and tracking techniques offer a wide variety of colors and functionalities to efficiently and sensitively identify and (or) follow specific cells within a population. Confocal microscopy offers several advantages over conventional optical microscopy, such as controllable depth of field, the elimination of image degrading information and collection of multiple sections from thick specimens. The high-quality images rendered by confocal microscopy from specimens prepared for conventional optical ones enable the information extracted from cell labeling and tracing to be more distinctive and accurate (Pawley, 2002) [89].

The laser-scanning cytometry (LSC) is a microscope-based cytofluorometer that has attributes of both flow and image cytometry. Individual cells are rapidly measured at multiple wavelengths with high accuracy and sensitivity. It is featured by carrying out cell tracking experiments that labeled cell samples can be examined repeatedly over time.
1.5 Background in modeling and simulation techniques

Complex biological systems are composed of a great number of entities that interact with each other to achieve system-level behaviors and processes. Computational approach for modeling and simulation devotes itself to tackle the inherent complexity by decomposing such systems into representative but simplified modules and components.

1.5.1 Event-based modeling and simulation

The event-based modeling captures the complex behavior of a system by denoting it with a set of time and space random variables. Thus, the entire system can be abstracted as a collection of functional modules and the system is driven by a sequence of events. The key underlying assumption is that the state space is decomposed into disjoint sets of independent events which can be executed simultaneously (Ghosh et al., 2007) [48].

Solving the event-based modeling often relies on the implementation of simulation particularly when it becomes infeasible to compute an exact result. The event-based simulation usually includes: (i) timescale - to keep track of the current simulation time in a suitable measurement unit; (ii) set of rules - to govern the state changing of each functional module and its actions in its environment; (iii) list of events - to keep a record of the start time and the end time of each event; (iv) random number generators - to generate different kinds of random variables; (v) statistics - to compute and save quantities that are of interest.
1.5.2 Monte Carlo simulation

Monte Carlo methods are often applied in the computational biology field to simulate biological systems that are abstracted by mathematical and physical entities. Its key algorithm is built on repeated random sampling, therefore using Monte Carlo simulation requires large amounts of random numbers.

Although different Monte Carlo methods may vary they all follow a general workflow pattern that is to: (i) firstly define a domain of possible inputs; (ii) secondly generate inputs randomly from probability distributions over the domain; (iii) thirdly perform a computation on the inputs; (iv) then repeatedly perform steps (ii) and (iii) for a large number of times; (v) and finally interpret the aggregated results.

1.5.3 Genetic algorithm

A genetic algorithm is a searching heuristic that mimics the process of natural evolution. It is used to generate optimized solutions to search problems in complex non-linear systems (Mitchell, 1996) [82].

We use genetic algorithm to carry out parameter search in the data fitting. Details about the algorithm used in this thesis are described as followed. Each parameter range is encoded by a bit vector with length $a$, yielding $2^a$ possible values. An initial pseudo “population” was created by setting $X$ randomly chosen parameter combinations as $X$ “individuals”. The value of each modeling parameter in any “individual” has been converted to binary format to become a 0-1 sequence. Each sequence can be treated as a “chromosome”. Thus, the “genome” of an “individual” consists of a complete “heritable” parameter setting. By evolving the “population”
under Wright-Fisher model for $Y$ generations with “mutation” and “crossover”, it yields by selection the “individuals” that can best fit the experimental observation. Normally each parameter range is binary coded as strings of 0s and 1s, but other encodings are also possible. We use binary coding throughout this dissertation.

1.5.4 simuPOP - a population genetics simulation environment

simuPOP is a general purpose individual-based forward time population genetics simulation environment (Peng and Kimmel, 2005). Its fundamental advantage is rooted in its simple and fast learned manner of allowing users to manipulate a powerful simulation tool to handle arbitrarily complex evolutionary scenarios. Although in terms of efficiency coalescent-based simulation methods may exceed forward-time approaches, it is inevitably superior to choose forward-time simulations if the interest is focused on the evolutionary process itself rather than solely its outcome (reviewed by Carvajal-Rodriguez, 2008), let alone that the impact of efficiency is becoming less and less important due to the development of fast-growing CPU speed, multi-thread computing technology and access to computer clusters.

Unlike other forward-time simulation tools, which either require users to adhere to predefined evolutionary processes with countable selections of demographic models and genetic factors, or ask users to implement scripts with tremendous efforts in a lower level programming language, such as FORTRAN/C/C++, simuPOP is currently the only one simulation program that provides both users’ simplicity of working on the Python programming platform and the flexibility of gluing pieces of modules and (or) functions to easily construct customized evolutionary scenarios.
at any level of complexity. Besides, built-in module simuOpt supports the need of graphical user interface in an easy and convenient way.

1.6 Outline of the dissertation

The dissertation is organized as follows. Chapter 2 introduces a stochastic model and simulation to study the ascertainment bias in discovery of microsatellite loci and explores factors that influence microsatellite allele size difference in the inter-species study. Chapter 3 brings out a stochastic simulation framework to generate and analyze sequence-based data for rare variant association studies, and evaluates association methods in an unbiased manner. Chapter 4 describes a mathematical model and stochastic simulation in analyzing cell population dynamics of the adult hippocampal neurogenesis. Chapter 5 presents a multiscale stochastic modeling and simulation of DNA replication in proliferating cell population, and shows that different rates of DNA replication at early versus late S-phase sections have impact on the experimental asymmetry in DNA-replication scatterplots.
Chapter 2

Ascertainment bias in discovery of microsatellite loci

2.1 Motivation and overview

Microsatellite loci play an important role as markers for identification, disease gene mapping and evolutionary studies. Mutation rate, which is of fundamental importance, can be obtained from interspecies comparisons, which however are subject to ascertainment bias. This bias arises for example when a locus is selected based on its large allele size in one species (cognate species 1), in which it is first discovered. This bias is reflected in average allele length in any non-cognate species 2 being smaller than that in species 1. This phenomenon was observed in various pairs of species, including comparisons of allele sizes in Human and Chimpanzee. Various mechanisms were proposed to explain observed differences in mean allele lengths between two species. In this chapter, we examine the framework of a single-step asymmetric and unrestricted stepwise mutation model with genetic drift. Analysis is based on coalescence theory. Analytical results are confirmed by simulations using the simuPOP software. The mechanism of ascertainment bias in this model is a tighter correlation of allele sizes within a cognate species 1 than of allele sizes in two
different species 1 and 2. We present computations of the expected average allele size difference, given the mutation rate, population sizes of species 1 and 2, time of separation of species 1 and 2, and the age of the allele. In particular, we show that when the past demographic histories of the cognate and non-cognate taxa are different, the rate and directionality of mutations will impact the allele sizes in the two taxa differently from the simple effect of ascertainment bias. This effect may exaggerate or reverse the effect of difference in mutation rates. We re-analyze literature data, which indicate that despite the bias, the microsatellite mutation rate estimate in the ancestral population is consistently greater than that in either Human or Chimpanzee and the mutation rate estimate in Human exceeds or equals to that in Chimpanzee with the rate of allele length expansion in Human being always greater than that in Chimpanzee. We also demonstrate that population bottlenecks and expansions in the recent human history have little impact on our conclusions.

2.2 Introduction

Ascertainment bias in population genetics is usually studied in two contexts. One of them is discovery of polymorphic loci and is best illustrated by the example of SNPs (single-nucleotide polymorphisms). Most of the published data on SNP sampling frequencies are obtained in a two-step process, where the first step involves discovering chromosomal locations of a number of SNPs, and the second one involves DNA sequencing of a sample of $n$ chromosomes restricted to locations discovered in the first step. The first step is called SNP discovery or ascertainment and is based on a number of chromosomes smaller than $n$. As demonstrated in a number
of papers, taking into account the ascertainment scheme is a very important aspect of SNP data analysis. For example, Polanski and Kimmel (2003) derived expressions for modeling the way in which ascertainment-modified SNP sampling frequencies and distorted inferences concerning the mutation rate. A more recent paper (Albrechtsen et al., 2010) considers chip-based high-throughput genotyping, which has facilitated genome-wide studies of genetic diversity. Many studies have utilized these large data sets to make inferences about the demographic history of human populations using measures of genetic differentiation such as $F_{ST}$ or principal component analysis. However, again, the single nucleotide polymorphism (SNP) chip data suffer from ascertainment biases caused by the SNP discovery process in which a small number of individuals from selected populations are used as discovery panels. Albrechtsen et al. (2010) generate SNP genotyping data for individuals that previously have been subject to partial genome-wide Sanger sequencing and compare inferences based on genotyping data to inferences based on direct sequencing. They demonstrate that the ascertainment bias distorts measures of human diversity and may change conclusions drawn from these measures in sometimes unexpected ways. They also show that details of the genotyping calling algorithms may have a surprisingly large effect on population genetic inferences. This type of ascertainment bias will be of importance in forthcoming genetic and genomic studies.

However, this chapter is concerned with a different type of ascertainment bias which occurs in inter-species or inter-population studies. If a genetic measure of variability or diversity such as heterozygosity, and its underlying causes such as
mutation, are studied in more than one species, a careful consideration of the sampling scheme used as basis for comparison is needed. Depending on from which species the polymorphisms are ascertained, the comparison of variability between the two species may be biased in a given direction. We will consider a specific scenario, in which two extant species, such as humans and chimpanzees, are traced to a common ancestral species. Also, we will consider microsatellite loci, which can be modeled mathematically in a relatively simple way, so that the forward-time simulations can be compared to analytical computations.

In our example, ascertainment bias of inter-species (population) studies of microsatellite loci occurs when a locus is selected based on its large allele size in the species in which it is first discovered (say, the cognate species 1). This bias is reflected in average allele length in any non-cognate species 2 being smaller than that in species 1. This phenomenon was observed in various pairs of species, including Human and Chimpanzee. Various mechanisms were proposed to explain the observed differences in mean allele lengths between two species. Here, we examine the simplest possible framework: A single-step asymmetric and unrestricted stepwise mutation model with genetic drift. The mathematical model is analyzed based on coalescence theory. The mechanism of ascertainment bias in this model is a tighter correlation of allele sizes within a cognate species 1 than of allele sizes in two different species 1 and 2. We present computations of the expected bias, given the mutation rate, population sizes of species 1 and 2, time of separation of species 1 and 2, and the age of the allele. In particular, using the coalescence theory, we show that when the demographic histories of the cognate and non-cognate species
are different, the rate and directionality of mutations impact the allele sizes in the two species differently than the simple effect of the ascertainment bias.

Microsatellite polymorphisms, characterized by variations of copy numbers of short motifs of nucleotides, have become a common tool for gene mapping and evolutionary studies since they are abundantly found in genomes of a large number of organisms (Pena *et al.*, 1993; Bowcock *et al.*, 1994; Deka *et al.*, 1994; Primmer and Ellegren, 1998) [90, 14, 31, 96]. High mutation rate at these loci is the attractive feature of using the microsatellites as tools for molecular evolutionary studies, since consequences of accumulation of past mutation events are seen as differences of allele frequency distributions even in closely related taxa (Weber and Wong, 1993; Kimmel and Chakraborty, 1996; Chakraborty *et al.*, 1997) [117, 66, 19]. However, in cross-species comparisons of allele size distributions at microsatellite loci some apparently discordant findings (namely, a systematic bias of average allele sizes in one species as compared to another), led some investigators to argue that these repeat loci may not be the most efficient tools for inter-species studies (Rubinsztein *et al.*, 1995; Crawford *et al.*, 1998) [103, 26]. In general, for evolutionary studies microsatellite loci as identified in one species (or population) are studied in other species (or populations), making use of their genome homology. Nevertheless, the process of detection (in the cognate species) and its use in a non-cognate species may inherently impact the allele size distribution and associated other summary measures of genetic variation (such as heterozygosity, allele size variance, or number of segregating alleles). This discordance, called the ascertainment bias, is claimed to have been observed in sheep (Forbes *et al.*, 1995) [46], swallows, cetaceans, ru-
minants, turtles and birds (Ellegren et al., 1995) [39]. However, Rubinsztein et al. (1995) [102] and Amos and Rubinsztein (1996) [6] explained such observations as inter-taxa differences of rates and patterns of mutations at microsatellite loci. The goal of the present study is to address this issue. Our approach is different from other attempts to study similar problems (see e.g. Rogers and Jorde, 1996 [100]). We consider a general model of mutations (called the Generalized Stepwise Mutation Model, GSMM) that is shown to be applicable to microsatellites (Kimmel et al., 1996; Kimmel and Chakraborty, 1996 [68, 66]), on which we superimpose the effects of demographic differences of cognate and non-cognate taxa, as both of these factors are known to jointly affect the features of polymorphisms at microsatellite loci in extant taxa (Kimmel et al., 1998) [67]. In particular, using the coalescence theory, we show that when the past demographic histories of the cognate and non-cognate taxa are different, the rate and directionality of mutations impact the allele sizes in the two taxa differently than the simple effect of ascertainment bias.

2.3 Modeling and simulation

2.3.1 Evolution of a DNA-repeat locus

We consider a DNA-repeat locus that has originated $t$ units of time ago (at backward or reverse time $t$), and observed at present (time 0). Adjective “backward” will be usually omitted. Chromosomes containing the locus belong to one of the two populations (labeled 1 and 2), which diverged $t_0$ time units before present (time $t_0$) from an ancestral population (labeled 0). The essentials are depicted in Figure 2.1.
Demographic scenario employed in the mathematical model and simuPOP simulations. Notation: $N_0$, $N_1$ and $N_2$, effective sizes of the ancestral, cognate and non-cognate populations, respectively; $X_0$, $X_1$ and $X_2$, increments of allele sizes due to mutations in the ancestral allele, in chromosome 1 and in chromosome 2, respectively.

The ancestral population consists of $2N_0$ chromosomes, and populations 1 and 2 of $2N_1$ and $2N_2$ chromosomes respectively. We assume the time-continuous Fisher-Wright-Moran model (Kimmel et al., 1998) [67]. At the locus considered, alleles mutate according to the unrestricted Generalized Stepwise Mutation Model (GSMM, Kimmel and Chakraborty, 1996) [66]. Specifically, the action of genetic drift and mutation can be represented by the following coalescence/mutation model:

1. Chromosomes 1 and 2, sampled at time 0 from populations 1 and 2, respectively, have a common ancestor $T$ units of time before present (Figure 2.1). Random variable $T$ has exponential distribution with parameter $1/(2N_0)$, shifted by $t_0$, i.e.

$$\Pr[T > \tau] = \begin{cases} 
1; & \tau \leq t_0, \\
\exp[-(\tau - t_0)/(2N_0)]; & \tau > t_0.
\end{cases} \quad (2.1)$$
In other words, as long as the two chromosomes or their direct ancestors belong to different populations (i.e. for $\tau \leq t_0$, in backward time), they cannot coalesce. From the moment the populations converge (i.e. for $\tau > t_0$ in reverse time), the distribution of the time to coalescence is exponential with parameter $1/(2N_0)$.

2. Chromosomes 1 and 1’, sampled at time 0 from population 1, have a common ancestor $T$ units of time before present, either in population 1, if $T \leq t_0$ or in the ancestral population 0, if $T > t_0$. Therefore, the random variable $T$ has a more complex distribution of the form,

$$
\Pr[T > \tau] = \begin{cases} 
\exp[-\tau/(2N_1)]; & \tau \leq t_0, \\
\exp[-t_0/(2N_1) - (\tau - t_0)/(2N_0)]; & \tau > t_0.
\end{cases}
$$

(2.2)

In other words, as long as the two chromosomes or their direct ancestors belong to population 1 (i.e. for $\tau \leq t_0$, in backward time), they coalesce with intensity $1/(2N_1)$. From the moment the species converge (i.e. for $\tau > t_0$ in backward time), the coalescence intensity is $1/(2N_0)$.

3. Initial size (number of repeats) at the locus at time ($t$) of the origin of the locus is equal to a constant. Choosing this constant equal to 0 is not a restrictive assumption. In our model, we assume that before time $t$ there were no mutation events.

4. Mutation epochs along the lines of descent occur according to a Poisson process with constant intensities $\nu_0$, $\nu_1$ and $\nu_2$ in populations 0, 1 and 2, respectively. Each mutation event alters the allele size $S$ by adding to it a random number
of repeats $U$, i.e.

$$S \rightarrow S + U.$$  

$U$ is an integer-valued random variable (rv) with probability generating function (pgf)

$$\varphi_k(s) = E(s^U) = \sum_{i=-\infty}^{\infty} \Pr[U = i]s^i.$$  

The pgf $\varphi_k(s)$ and, equivalently, the distribution of $U$ is generally different in each population $k$ ($k = 0, 1, 2$). Consequently, the change of the allele size, during a time interval of length $\Delta t$ spent in population $k$ is a compound Poisson random variable with pgf $\exp\{\nu \Delta t[\varphi_k(s) - 1]\}$. For the asymmetric single-step Stepwise Mutation Model (SSMM), we have

$$\varphi_k(s) = b_k s + d_k / s,$$  \hspace{1cm} (2.3)

where $b_k = \Pr[U = 1]$ and $d_k = \Pr[U = -1] = 1 - b_k$ are the respective probabilities of expansion and contraction of the allele in a single mutation epoch.

Remark. The model is formulated as if the length of generation in species 0, 1, and 2 were identical. However, the mutation rates and populations sizes can be rescaled, to accommodate different generation time as explained in the section concerning modeling (see further on). Indeed all results in the following section are invariant under rescaling. We will return to this issue in the discussion.
2.3.2 Conditional distributions and ascertainment bias of allele sizes

The main purpose of this section is to use the principles of the coalescence theory (as reviewed by Tavaré, 1984 [111]) to derive conditional expected allele size at a chromosome, given the allele size on another chromosome sampled either from a different or from the same population as the original chromosome. This information is crucial for obtaining theoretical estimates of the ascertainment bias in conjunction with other effects.

2.3.2.1 Chromosomes sampled from populations 1 and 2

We use notation as in Figure 2.1: \( X_0, X_1 \) and \( X_2 \) denote the incremental changes of allele sizes in the ancestral chromosome, 0, of chromosomes 1 and 2, and in chromosomes 1 and 2, respectively. Conditionally on \( T \), \( X_0, X_1 \) and \( X_2 \) are independent random variables. Let us note that while chromosome 0 always lives in population 0, chromosomes 1 and 2 begin their lives in population 0 and then continue in populations 1 and 2. Let \( Y_1 = X_0 + X_1 \) and \( Y_2 = X_0 + X_2 \) denote the allele sizes at time 0 (present time) at chromosomes 1 and 2 respectively. We want to find the expected allele size at chromosome 2, jointly with the allele size at chromosome 1 being equal to \( i \) (conditional on \( \{ T = \tau \} \)),

\[
E[Y_2; Y_1 = i | T = \tau] = \sum_j E[X_0 + X_2; X_0 = j; X_1 = i - j | T = \tau]
\]

\[
= E[X_2 | T = \tau] \Pr[Y_1 = i | T = \tau]
\]

\[
+ \sum_j j \Pr[X_0 = j | T = \tau] \Pr[X_1 = i - j | T = \tau]. \tag{2.4}
\]
If we translate the above into the language of probability generating functions, we obtain

\[
\sum_i E[Y_2; Y_1 = i | T = \tau] s^i = E[X_2 | T = \tau] f_{X_0 | T = \tau}(s) f_{X_1 | T = \tau}(s) + s f'_{X_0 | T = \tau}(s) f_{X_1 | T = \tau}(s).
\]

(2.5)

See section 2.3.4 for more details about probability generating function and how to derive equation (2.5) from equation (2.4).

2.3.2.2 Chromosomes sampled from population 1

Using the same reasoning as above, we obtain

\[
\sum_i E[Y'_1; Y_1 = i | T = \tau] s^i = E[X'_1 | T = \tau] f_{X_0 | T = \tau}(s) f_{X_1 | T = \tau}(s) + s f'_{X_0 | T = \tau}(s) f_{X_1 | T = \tau}(s).
\]

(2.6)

2.3.2.3 Probability generating functions and expectations of incremental changes of allele sizes

Random variables \(X_0, X_1\) and \(X_2\) result from compounding the Poisson process (Kingman, 1993) of mutations, with varying intensities \(\nu_0, \nu_1\) and \(\nu_2\), by matching distributions of allele size changes with pgf’s \(\varphi_0(s), \varphi_1(s)\) and \(\varphi_2(s)\) respectively. The choice of intensity and pgf depends on the population in which the chromosomes reside during a given time interval. Without getting into detail, we obtain:

\[
f_{X_0 | T = \tau}(s) = \begin{cases} 
\exp\{ (t - t_0)\nu_0[\varphi_0(s) - 1] + (t_0 - \tau)\nu_1[\varphi_1(s) - 1] \}; & \tau \leq t_0, \\
\exp\{ (t - \tau)\nu_0[\varphi_0(s) - 1] \}; & t_0 < \tau \leq t, \\
1; & \tau > t, 
\end{cases}
\]

(2.7)
\[
f_{X_i|T=\tau}(s) = \begin{cases} 
\exp\{\tau \nu_i [\varphi_i(s) - 1]\}; & \tau \leq t_0, \\
\exp\{(\tau - t_0) \nu_0 [\varphi_0(s) - 1] + t_0 \nu_i [\varphi_i(s) - 1]\}; & t_0 < \tau \leq t, \\
\exp\{(t - t_0) \nu_0 [\varphi_0(s) - 1] + t_0 \nu_i [\varphi_i(s) - 1]\} & \tau > t,
\end{cases}
\] (2.8)

for \(i = 1, 2\). Also, \(f_{X_1'|T=\tau}(s) \equiv f_{X_1|T=\tau}(s)\). The conditional expected values are obtained by differentiation of respective pgf’s and setting \(s = 1\).

### 2.3.3 Mean allele size difference, \(D\)

In the single-step stepwise mutation model (SSMM), the pgf’s \(\varphi_0(s)\), \(\varphi_1(s)\) and \(\varphi_2(s)\) have the form as in Equ. (2.3). We note the expansion

\[
e^{\nu t [bs + d/s - 1]} = \sum_{i \in \mathbb{Z}} \beta_i s^i = \sum_{i \in \mathbb{Z}} e^{-\nu t} I_i (2 \nu t \sqrt{bd}) \left( \frac{b}{d} \right)^{i/2} s^i,
\] (2.9)

valid for \(|s| = 1\), where \(I_i = I_{-i}\) is the modified Bessel function of the first type, of integer order \(i\) (Refs. Abramowitz and Stegun, 1972 [2]). Using this expansion, it is possible to represent the right-hand sides of Eqs. (2.5) and (2.6) as power series in variable \(s\). Finally,

\[
E[Y_2; Y_1 = i] = \int_0^\infty E[Y_2; Y_1 = i|T = \tau] f_T(\tau) d\tau,
\] (2.10)

\[
E[Y_1'; Y_1 = i] = \int_0^\infty E[Y_1'; Y_1 = i|T = \tau] f_T(\tau) d\tau,
\] (2.11)

where \(f_T(\tau)\) is the distribution density of the time to coalescence, based on relationships (2.1) and (2.2) respectively. A computational expression for \(Pr[Y_1 = i]\)
can be similarly obtained from

\[ \Pr[Y_1 = i] = \int_0^\infty \Pr[Y_1 = i|T = \tau]f_T(\tau)d\tau \]  \hspace{1cm} (2.12)

Suppose that a DNA-repeat locus discovered in a genome search of population 1 is retained for further study if it has a minimum number of \( x \) repeats of the motif, i.e. if

\[ Y_1 \geq x. \]

This criterion is also a substitute measure of this locus’s variability, and hence of its polymorphism. The reason is that, irrespective of directionality of mutational changes, in the GSMM, the extremes of repeat count are strongly positively correlated with variance of repeat count and heterozygosity at the locus. This latter is a consequence of the random-walk mechanism of mutations in this model (for a discussion and references, see e.g. Kimmel and Chakraborty, 1996 [66]).

If the locus is retained and a sample of \( n \) individuals from the non-cognate population 2 is typed for this locus, then the expected value of the mean repeat count in the sample is equal to

\[ E\left(\frac{1}{n} \sum_{i=1}^{n} Y_{2i} | Y_1 \geq x\right) = E[Y_2 | Y_1 \geq x] = \frac{\sum_{i \geq x} E[Y_2; Y_1 = i]}{\sum_{i \geq x} \Pr[Y_1 = i]} \]  \hspace{1cm} (2.13)

If a sample of \( n \) individuals of the cognate population 1 is typed for this locus, then the expected values of the mean repeat count in the sample is equal to

\[ E\left(\frac{1}{n} \sum_{i=1}^{n} Y_{1i} | Y_1 \geq x\right) = E[Y_1 | Y_1 \geq x] = \frac{\sum_{i \geq x} E[Y_1'; Y_1 = i]}{\sum_{i \geq x} \Pr[Y_1 = i]} \]  \hspace{1cm} (2.14)
The mean allele size difference, $D$, which is due to a combined effect of ascertainment bias and intrinsic genetic factors, can be defined as

$$D = E[Y_1'|Y_1 \geq x] - E[Y_2|Y_1 \geq x]. \quad (2.15)$$

### 2.3.4 Derivation of the analytical solution to $D$

The analytical solution of equation (2.15) can be obtained from derivations of equations shown below.

#### 2.3.4.1 Derivation of equation (2.4)

Since $Y_2 = X_0 + X_2$ and $Y_1 = X_0 + X_1$, we have

$$E[Y_2; Y_1] = i|T = \tau]$$

$$= E[X_0 + X_2; X_0 + X_1 = i|T = \tau]$$

$$= \sum_j E[X_0 + X_2; X_0 = j; X_1 = i - j|T = \tau]$$

$$= \sum_j E[X_0; X_0 = j; X_1 = i - j|T = \tau] + \sum_j E[X_2; X_0 = j; X_1 = i - j|T = \tau]$$

$$= \sum_j jP(X_0 = j, X_1 = i - j|T = \tau)$$

$$+ \sum_j E[X_2|T = \tau]P(X_0 = j, X_1 = i - j|T = \tau)$$

$$= \sum_j jP(X_0 = j|T = \tau)P(X_1 = i - j|T = \tau)$$

$$+ E[X_2|T = \tau]\sum_j P(X_0 = j, X_1 = i - j|T = \tau)$$

$$= \sum_j jP(X_0 = j|T = \tau)P(X_1 = i - j|T = \tau)$$

$$+ E[X_2|T = \tau]P(X_0 + X_1 = i|T = \tau)$$

$$= \sum_j jP(X_0 = j|T = \tau)P(X_1 = i - j|T = \tau)$$

$$+ E[X_2|T = \tau]P(X_0 + X_1 = i|T = \tau)$$
2.3.4.2 Derivation of equations (2.5) and (2.6)

If $X$ is a discrete random variable taking values in the non-negative integers $\{0, 1, 2, \ldots\}$, then we define the probability generating function (pgf) of $X$ as:

$$f_X(s) = E[s^X] = \sum_{x=0}^{\infty} P(X = x)s^x$$

Restating equation (2.4) in the terms of the probability generating functions, and using independence of $(X_0|T), (X_1|T), (X_2|T)$, we obtain

$$\sum_i \sum_j E[Y_{2,i}Y_{1,j}] = \sum_i E[X_2|T = \tau]P(X_0 + X_1 = i|T = \tau)s^i$$

$$= \sum_i E[X_2|T = \tau]P(X_0 + X_1 = i|T = \tau)s^i$$

$$+ \sum_j \sum_i jP(X_0 = j|T = \tau)P(X_1 = i - j|T = \tau)s^{i-j}$$

$$= E[X_2|T = \tau]f_{X_0+X_1|T=\tau}(s) + \sum_i f'_{X_0|T=\tau}(s)\sum_{i=\infty}^{\infty} P(X_1 = i - j|T = \tau)s^{i-j}$$

$$= E[X_2|T = \tau]f_{X_0+X_1|T=\tau}(s) + \sum_{i=\infty}^{\infty} f'_{X_0|T=\tau}(s)\sum_{i-j=-\infty}^{\infty} P(X_1 = i - j|T = \tau)s^{i-j}$$

(since $-\infty < j < \infty$)

$$= E[X_2|T = \tau]f_{X_0|T=\tau}(s)f_{X_1|T=\tau}(s) + sf'_{X_0|T=\tau}(s)f_{X_1|T=\tau}(s)$$

Thus, equation (2.5) holds; derivation of equation (2.6) is similar.

2.3.4.3 Derivation of equations (2.7) and (2.8)

Consider the Poisson Process $N$ with intensity $\nu$, where $N(T) = \# \{\text{events of mutations occurring in time interval of length } T\}$. 
\( f_{N(T)}(s) = e^{\nu t(s-1)} \) and \( T \) is the coalescence time before present. Let \( X_0, X_1, X_2 \) denote the incremental changes of allele length of chromosomes 0, 1 and 2. \( (X_0|T), (X_1|T), (X_2|T) \) are conditionally independent and \( f_{U_k}(s) = \psi_k(s), k = 0, 1, 2. \)

Therefore, if \( \tau \leq t_0 \), it holds

\[
X_0|(T = \tau) = X_{0,1}|(T = \tau) + X_{0,2}|(T = \tau), \text{ where } X_{0,1}|T = \tau \text{ is the increment of allele size in the interval } [t, t_0] \text{ and } X_{0,2}|T = \tau \text{ is the increment in the interval } [t_0, \tau].
\]

Since \( X_{0,1}|(T = \tau) \) is independent of \( X_{0,2}|(T = \tau) \), it holds that \( \text{pgf} \ f_{X_0|T=\tau}(s) = f_{X_{0,1}|T=\tau}(s) \cdot f_{X_{0,2}|T=\tau}(s) \), where \( f_{X_{0,1}|T=\tau}(s) = f_{N(t-t_0)}(\psi_0(s)) = e^{\nu(t-t_0)(\psi_0(s)-1)}, \)
and \( f_{X_{0,2}|T=\tau}(s) = f_{N(t_0-\tau)}(\psi_1(s)) = e^{\nu_1(t_0-\tau)(\psi_1(s)-1)}. \)

It follows that \( f_{X_0|T=\tau}(s) = \exp(\nu_0(t-t_0)(\psi_0(s)-1) + \nu_2(t_0-\tau)(\psi_1(s)-1)), \) and \( f_{X_1|T=\tau}(s) = f_{N(\tau)}(\psi_1(s)) = \exp(\nu_1(\psi_1(s)-1)). \) Therefore, both equations (2.7) and (2.8) hold for \( \tau \leq t_0. \) Derivations of \( t_0 < \tau < t \) or \( \tau > t \) are similar.

2.3.4.4 Derivation of computational expressions for equations (2.10) and (2.11)

Expected size of allele drawn from population 2, jointly with size of allele drawn from population 1 being equal to \( i \) is equal to \( E[Y_2, Y_1 = i] = \int_0^\infty E[y_2, Y_1 = i|T = \tau] f_T(\tau) d\tau, \) where \( f_T(\tau) \) is given in equation (2.1). However, \( T \) denotes the common ancestor time of chromosomes 1 and 2 sampled at time 0 from populations 1 and 2. Therefore, from Equ. (2.1), we have

\[
f_T(\tau) = \begin{cases} 
0 & \tau \leq t_0 \\
\frac{1}{2N_0} e^{-(\tau - t_0)/(2N_0)} & \tau > t_0 
\end{cases}
\]

If \( \tau \leq t_0, \) \( f_T(\tau) = 0 \) and \( E[Y_2; Y_1 = i] = 0. \)
If \( t_0 < \tau \leq t \), from Equ. \((2.5)\) we obtain 
\[
\sum_i E[Y_2, Y_1 = i| T = \tau] s^i = E[X_2|T = \tau] f_{X_0|T = \tau}(s) f_{X_1|T = \tau}(s) + s f'_{X_0|T = \tau}(s) f_{X_1|T = \tau}(s)
\]
where,

\[
E[X_2|T = \tau] f_{X_0|T = \tau}(s) f_{X_1|T = \tau}(s) = E[X_2|T = \tau] \exp(\nu_0(t - t_0)(\psi_0(s) - 1) + \nu_1 t_0 (\psi_1(s) - 1))
\]
\[
= E[X_2|T = \tau] \exp(\nu_0(t - t_0) (b_0 s + \frac{d_0}{s} - 1) + \nu_1 t_0 (b_1 s + \frac{d_1}{s} - 1))
\]
\[
= E[X_2|T = \tau] \exp((\nu_0(t - t_0)b_0 + \nu_1 t_0 b_1)s + (\nu_0(t - t_0)d_0 + \nu_1 t_0 d_1) / s - \nu_0(t - t_0) - \nu_1 t_0)
\]
\[
E[X_2|T = \tau] = \tau e^{(\nu_0(t - t_0)b_0 + \nu_1 t_0 b_1) s + (\nu_0(t - t_0)d_0 + \nu_1 t_0 d_1) / s} e^{-\nu_0(t - t_0) - \nu_1 t_0}
\]

\((t_0 < \tau \leq t)\)

By differentiating \( f_{X_2|T = \tau}(s) \) and setting \( s = 1 \), we obtain

\[
E[X_2|T = \tau] = (\nu_0(\tau - t_0)(b_0 - d_0) + \nu_2 t_0 (b_2 - d_2)) \exp(\nu_0(\tau - t_0)(b_0 + d_0 - 1) + \nu_2 t_0 (b_2 + d_2 - 1)) (t_0 < \tau \leq t). \]

We denote \( b = \nu_0(t - t_0) b_0 + \nu_1 t_0 b_1 \) and \( d = \nu_0(t - t_0) d_0 + \nu_1 t_0 d_1 \),

to obtain \( e^{((\nu_0(t - t_0) b_0 + \nu_1 t_0 b_1) s + (\nu_0(t - t_0) d_0 + \nu_1 t_0 d_1) / s)} = e^{(b s + d) / s} \). According to Equ. \((2.9)\),

\[
e^{(b s + d) / s} = \sum_{i \in \mathbb{Z}} I_i (2 \sqrt{bd}) (\frac{b}{d})^{i/2} s^i = \sum_{i \in \mathbb{Z}} \beta_i s^i , \ |s| = 1,
\]

where we denote \( \beta_i = I_i (2 \sqrt{bd}) (\frac{b}{d})^{i/2} \) , and \( I_i = I_{-i} \) is the modified Bessel function of the first type (Abramowitz and Stegun, 1972) [2], of integer order \( i \)

41
Thus

\[ E[X_2|T = \tau]f_{X_0|T=\tau}(s)f_{X_1|T=\tau}(s) \]
\[ = e^{-\nu_0(t-t_0)-\nu_1 t_0} \sum_{i \in \mathbb{Z}} E[X_2|T = \tau] \beta_i s^i \]

Furthermore, \( sf'_{X_0|T=\tau}(s)f_{X_1|T=\tau}(s) \) in equation (2.5) can be expressed as

\[
 sf'_{X_0|T=\tau}(s)f_{X_1|T=\tau}(s) = sf'_{X_0|T=\tau}(s)f_{X_1|T=\tau}(s) \\
 = s\nu_0(t - \tau)\psi'_0(s) \times e(\nu_0(t - t_0)(\psi_0(s) - 1) + \nu_1t_0(\psi_1(s) - 1)) \\
 = s\nu_0(t - \tau)(b_0 - d_0/s^2) \\
 \times e(\nu_0(t-t_0)b_0 + \nu_1t_0b_1 + (\nu_0(t-t_0)d_0 + \nu_1t_0d_1)/s) e^{-\nu_0(t-t_0)-\nu_1 t_0} \\
 = \nu_0(t - \tau)(b_0 s - d_0/s) e^{-\nu_0(t-t_0)-\nu_1 t_0} (\sum_{i \in \mathbb{Z}} \beta_i s^i) \\
 = e^{-\nu_0(t-t_0)-\nu_1 t_0} (\sum_{i \in \mathbb{Z}} \nu_0(t - \tau)b_0 \beta_i s^{i+1} - \sum_{i \in \mathbb{Z}} \nu_0(t - \tau)d_0 \beta_i s^{i-1}) \\
 = e^{-\nu_0(t-t_0)-\nu_1 t_0} (\nu_0(t - \tau)b_0 \sum_{i \in \mathbb{Z}} \beta_{i-1} s^i - \nu_0(t - \tau)d_0 \sum_{i \in \mathbb{Z}} \beta_{i+1} s^i) \\

Therefore, when \( t_0 < \tau \leq t \)

\[
 \sum_i E[Y_2, Y_1 = i|T = \tau]s^i = E[X_2|T = \tau]f_{X_0|T=\tau}(s)f_{X_1|T=\tau}(s) + sf'_{X_0|T=\tau}(s)f_{X_1|T=\tau}(s) \\
 = e^{-\nu_0(t-t_0)-\nu_1 t_0} \sum_{i \in \mathbb{Z}} [E[X_2|T = \tau] \beta_i \\
 + \nu_0(t - \tau)b_0 \beta_{i-1} - \nu_0(t - \tau)d_0 \beta_{i+1}]s^i \\
 = e^{-\nu_0(t-t_0)-\nu_1 t_0} \sum_{i \in \mathbb{Z}} \{\nu_0(t - \tau)b_0 \beta_{i-1} - \nu_0(t - \tau)d_0 \beta_{i+1} \\
 + [\nu_0(\tau - t_0)(b_0 - d_0) + \nu_2t_0(b_2 - d_2)] \}
 \times e^{\nu_0(t-t_0)(b_0+d_0-1)+\nu_2t_0(b_2+d_2-1)\beta_1} s^i \\

42
which yields

\[ E[Y_2, Y_1 = i|T = \tau] = e^{-\nu_0(t-t_0)-\nu_1t_0} \{ \nu_0(t - \tau)b_0\beta_{i-1} - \nu_0(t - \tau)d_0\beta_{i+1} \\
+ [\nu_0(\tau - t_0)(b_0 - d_0) + \nu_2t_0(b_2 - d_2)]e^{\nu_0(\tau-t_0)(b_0+d_0-1)+\nu_2t_0(b_2+d_2-1)\beta_i} \}
\]

which provides the desired computable form of Equ. (2.10).

If \( \tau > t \), analogous computations yield

\[
E[Y_2, Y_1 = i] = \int_t^\infty E[Y_2, Y_1 = i|T = \tau]f_T(\tau)d\tau
= E[Y_2, Y_1 = i|T = \tau]P[T > \tau]
= e^{-\frac{\nu_0t_0}{2\nu_0} - \nu_1t_0 - \nu_0(t-t_0)}\beta_i(\nu_0(t - t_0)(b_0 - d_0) + \\
+ \nu_2t_0(b_2 - d_2))e^{\nu_0(t-t_0)(b_0+d_0-1)+\nu_2t_0(b_2+d_2-1)}
\]

Similarly as derivation of Equ. (2.10), Equ. (2.11) can be fully derived from Equ. (2.2), (2.6) and (2.9) for its computable form, where

if \( \tau \leq t_0 \),

\[
E[Y'_1, Y_1 = i|T = \tau] = e^{-\nu_0(t-t_0)-\nu_1t_0} \{ [\nu_0(t - t_0)b_0 + \nu_1(t_0 - \tau)b_1]\beta_{i-1} + \nu_1\tau(b_1 - d_1)e^{\nu_1\tau(t_1+d_1-1)\beta_i} - [\nu_0(t - t_0)d_0 + \nu_1(t_0 - \tau)d_1]\beta_{i+1} \}
\]

if \( t_0 < \tau \leq t \),

\[
E[Y'_1, Y_1 = i|T = \tau] = e^{-\nu_0(t-t_0)-\nu_1t_0} \{ \nu_0(t - \tau)b_0\beta_{i-1} - \nu_0(t - \tau)d_0\beta_{i+1} + [\nu_0(\tau - t_0)(b_0 - d_0) + \nu_1t_0(b_1 - d_1)]e^{\nu_0(\tau-t_0)(b_0+d_0-1)+\nu_1t_0(b_1+d_1-1)\beta_i} \}
\]

if \( \tau > t \),

\[
E[Y'_1, Y_1 = i|T = \tau] = e^{-\nu_0(t-t_0)-\nu_1t_0}[\nu_0(t-t_0)(b_0-d_0)+\nu_1t_0(b_1-d_1)]e^{\nu_0(t-t_0)(b_0+d_0-1)+\nu_1t_0(b_1+d_1-1)\beta_i}
\]
2.3.4.5 Derivation of computational expression for equation (2.12)

The probability generating function of \((Y_1|T) = (X_0|T) + (X_1|T)\) is equal to 
\(f_{X_0|T=\tau}(s)f_{X_1|T=\tau}(s)\) because \((X_0|T)\) is conditionally independent of \((X_1|T)\).

Therefore,

\[
\sum_i P(Y_1 = i | T = \tau) s^i = f_{X_0|T=\tau}(s)f_{X_1|T=\tau}(s) = \exp(\nu_0(t - t_0)(\psi_0(s) - 1) + \nu_1 t_0(\psi_1(s) - 1)) = e^{-\nu_0(t - t_0) - \nu_1 t_0} \sum_{i \in \mathbb{Z}} \beta_i s^i
\]

which yields \(P(Y_1 = i) = \int_0^\infty P(Y_1 = i | T = \tau) f_T(\tau) d\tau = e^{-\nu_0(t - t_0) - \nu_1 t_0} \beta_i\), the derived computational expression for Equ. (2.12).

2.3.4.6 Derivation of equations (2.13) and (2.14)

By definition,

\[
E[Y_2|Y_1 \geq x] = \frac{E[Y_2,Y_1 \geq x]}{P(Y_1 \geq x)} = \frac{\sum_j j P(Y_2 = j,Y_1 \geq x)}{\sum_{i \geq x} P(Y_1 = i)} = \frac{\sum_j j \sum_{i \geq x} P(Y_2 = j,Y_1 = i)}{\sum_{i \geq x} P(Y_1 = i)} = \frac{\sum_{i \geq x} (\sum_j j P(Y_2 = j,Y_1 = i))}{\sum_{i \geq x} P(Y_1 = i)} = \frac{\sum_{i \geq x} E[Y_2|Y_1 = i]}{\sum_{i \geq x} P(Y_1 = i)}
\]

Equ. (2.13) has been derived. Similarly, Equ. (2.14) can be derived.
2.3.5 Simulation method

Despite the complexity of the theory involved in the study of ascertainment bias, simulation of such a process is straightforward using simuPOP (Peng and Kimmel, 2005) [91] (see detail about simuPOP in section 1.5.4). We consider a microsatellite locus founder population with \( N_0 \) individuals (\( 2N_0 \) chromosomes). We consider a diploid with initial allele of size on each chromosome 100. The founder population is evolved for \( t - t_0 \) generations before two copies of this population of sizes \( N_1 \) and \( N_2 \) were created, which are evolved for another \( t_0 \) generations. Direct execution of simulations for tens of thousands of generations is time-consuming. The probability that a random allele exceeds a specified threshold may be low, therefore many attempts may be needed to obtain an estimate of ascertainment bias. See Figure 2.2 for an illustrated workflow using simuPOP.

This problem can be addressed through the use of a scaling technique (Hoggart et al., 2007) [56]. Compared to a regular simulation that evolves a population
of size $N$ for $t$ generations, a scaled simulation with a scaling factor $\lambda$ evolves a smaller population of size $N/\lambda$ for $t/\lambda$ generations with magnified (multiplied by $\lambda$) mutation, recombination, and selection forces. This method can be justified by a diffusion approximation to the standard Wright-Fisher process (Ewens, 2004; Hoggart et al., 2007) \(^43\) \(^56\); however, because the diffusion approximation only applies to weak genetic forces in the evolution of haploid sequences, it cannot be involved when nonadditive diploid or strong genetic forces are used. Simulation study has been performed with a scaling factor $\lambda$, where populations with sizes $N_i/\lambda$ are evolved for $t_i/\lambda$ generations, under mutation models with mutation rates $\lambda \nu_i$, where $N_i \sim 10^4 - 10^6$, $t_i \sim 10^3 - 10^5$ and $\nu_i \sim 10^{-4}$ are values typical of human and primate effective population sizes, evolutionary history and microsatellite mutation rates. Running the simulations with different scaling factors yields identical results if $\lambda \leq 100$ ($\lambda = 1000, 500, 100, 50, 10$ have been tried).

\section*{2.4 Results}

\subsection*{2.4.1 Summary of modeling results}

The purpose of modeling is to determine in what circumstances the presence or absence of differences, observed in sizes of alleles at loci discovered in a cognate species (population 1) and then typed in a non-cognate species (population 2), can be attributed to ascertainment bias, or alternatively to differential effects of genetic drift or mutation rate and pattern. Before we present the numerical result, let us review the intuitions concerning these effects. These intuitions are valid independently of a particular model of mutations:
1. The observed difference between allele sizes per se, Eq. (2.15), results from a stronger correlation between allele states of chromosomes in cognate population 1 as compared to non-cognate population 2.

2. Reduced genetic drift in population 1 may reduce the effects of ascertainment bias. Indeed, if the cognate population 1 is much larger than the non-cognate population 2, then the coalescence process within population 1 has the star-like structure characterized by reduced dependence of allele states (Tajima, 1989) \[110\]. Therefore, the difference between correlations of allele states of chromosomes in cognate population 1 and non-cognate population 2, will be reduced. Note that the size of the non-cognate population 2 will not influence the difference of expected allele sizes, but it may influence other indices of polymorphism.

3. Mutation rate and pattern, different in populations 1 and 2, influence the differences in allele sizes between different populations.

Figure 2.3 depicts a series of modeling studies of \(D\), the combined effect of ascertainment bias, genetic drift and differential mutation rate on the mean repeat count, based on simuPOP model, compared to those obtained using Eq. (2.15). Parameter values approximate the evolutionary dynamics of dinucleotides in humans and chimpanzees: time from divergence of species \(t_0 = 4 \times 10^6\) years = \(2 \times 10^5\) generations for Figure 2.3 (assuming 20 years per generation), the age of the repeat locus \(t = 1 \times 10^7\) years = \(5 \times 10^5\) generations, mutation rate \(\nu = 1 \times 10^{-4}\) per generation, and probability of increase of allele size in a single mutation event, \(b = 0.55\).

Figure 2.3A depicts the values of \(D\) for the basic parameter values \(b_0 = b_1 = b_2 =\)
Figure 2.3: Observed difference $D$ in allele sizes may be positive or negative

Comparison of SimuPOP simulations with computations using mathematical expressions based on Equ. (2.15). (A) Values of $D$ for the basic parameter values $b_0 = b_1 = b_2 = b = 0.55$, $\nu_0 = \nu_1 = \nu = 0.0001$, $t_0 = 2 \times 10^5$ generations and $t = 5 \times 10^5$ generations, with the effective sizes of all populations concurrently varying from $2 \times 10^4$ to $4 \times 10^5$ individuals and with mutation rates $\nu_2$ varying from $\nu$ to $5\nu$. (B) Values of $D$ for the basic parameter values $b_0 = b_1 = b_2 = b = 0.55$, $\nu_0 = \nu_2 = \nu = 0.0001$, $t_0 = 2 \times 10^5$ generations and $t = 5 \times 10^5$ generations, with the effective sizes of all populations concurrently varying from $2 \times 10^4$ to $4 \times 10^5$ individuals and with mutation rates $\nu_1$ varying from $\nu$ to $5\nu$ (assuming 20 years per generation).
$b = 0.55$, and $\nu_0 = \nu_1 = \nu = 0.0001$, with the effective sizes of all populations concurrently varying from $2 \times 10^4$ to $4 \times 10^5$ individuals and with mutation rates $\nu_2$ varying from $\nu$ to $5\nu$. Figure 2.3B depicts the values of $D$ for the basic parameter values $b_0 = b_1 = b_2 = b = 0.55$, and $\nu_0 = \nu_2 = \nu = 0.0001$, with the effective sizes of all populations concurrently varying from $2 \times 10^4$ to $4 \times 10^5$ individuals and with mutation rates $\nu_1$ varying from $\nu$ to $5\nu$. These two figures make it explicit that the combined effect of ascertainment bias, genetic drift and differential mutation rate on the mean repeat count can result in a range of $D$ values from positive to negative ones.

### 2.4.2 Genetic algorithm and plausible parameter ranges

For the purpose of obtaining sets of model parameters that yield good fit to the experimental observation of allele length differences, we have applied the genetic algorithm (Mitchell, 1996) [82] as a search heuristic to explore an arguably realistic parameter space that specifies a variety of discrete values within a reasonable range to each of the key parameters; see section 1.5.3 for detail about genetic algorithm. We set $t$ to vary in the range from 440,000 to 740,000; $t_0$ from 250,000 to 400,000; $N_0$ from 10,000 to 85,000; $N_1$ from 5,000 to 12,000; $N_2$ from 10,000 to 25,000; $\nu_0, \nu_1, \nu_2$ from $5 \times 10^{-5}$ to $1 \times 10^{-3}$; $b_0, b_1, b_2$ from .51 to .55; $x$ from 12 to 18. Section 2.5 (Discussions and conclusions) involves more detail about settings of these ranges. We compare modeling results to observations of Cooper et al., 1998 [25]; see the next section for detail.
Table 2.1: Parameter settings that yield a good fit, for a range of realistic effective population sizes and mutation rates

<table>
<thead>
<tr>
<th></th>
<th>t</th>
<th>t₀</th>
<th>N₀</th>
<th>N₁</th>
<th>N₂</th>
<th>ν₀</th>
<th>ν₁</th>
<th>ν₂</th>
<th>b₀</th>
<th>b₁</th>
<th>b₂</th>
<th>x</th>
<th>D_HC</th>
<th>D_CH</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>540</td>
<td>270</td>
<td>15</td>
<td>6</td>
<td>15</td>
<td>.00012</td>
<td>.00006</td>
<td>.00002</td>
<td>.55</td>
<td>.55</td>
<td>.55</td>
<td>12</td>
<td>5.11</td>
<td>1.30</td>
</tr>
<tr>
<td></td>
<td>550</td>
<td>280</td>
<td>10</td>
<td>9</td>
<td>10</td>
<td>.00012</td>
<td>.00006</td>
<td>.00002</td>
<td>.55</td>
<td>.55</td>
<td>.55</td>
<td>12</td>
<td>5.06</td>
<td>1.10</td>
</tr>
<tr>
<td></td>
<td>570</td>
<td>300</td>
<td>15</td>
<td>3</td>
<td>20</td>
<td>.00030</td>
<td>.00022</td>
<td>.00016</td>
<td>.55</td>
<td>.55</td>
<td>.55</td>
<td>12</td>
<td>5.35</td>
<td>1.29</td>
</tr>
<tr>
<td></td>
<td>560</td>
<td>290</td>
<td>20</td>
<td>5</td>
<td>20</td>
<td>.00030</td>
<td>.00022</td>
<td>.00016</td>
<td>.55</td>
<td>.55</td>
<td>.55</td>
<td>12</td>
<td>5.27</td>
<td>1.14</td>
</tr>
<tr>
<td></td>
<td>460</td>
<td>250</td>
<td>25</td>
<td>10</td>
<td>12</td>
<td>.00030</td>
<td>.00055</td>
<td>.00045</td>
<td>.55</td>
<td>.55</td>
<td>.55</td>
<td>12</td>
<td>5.42</td>
<td>1.24</td>
</tr>
<tr>
<td>B</td>
<td>580</td>
<td>260</td>
<td>10</td>
<td>7</td>
<td>25</td>
<td>.00075</td>
<td>.00020</td>
<td>.00005</td>
<td>.51</td>
<td>.51</td>
<td>.51</td>
<td>18</td>
<td>5.18</td>
<td>1.26</td>
</tr>
<tr>
<td></td>
<td>720</td>
<td>250</td>
<td>15</td>
<td>8</td>
<td>23</td>
<td>.00015</td>
<td>.00010</td>
<td>.00005</td>
<td>.55</td>
<td>.55</td>
<td>.55</td>
<td>18</td>
<td>5.17</td>
<td>1.22</td>
</tr>
<tr>
<td></td>
<td>620</td>
<td>250</td>
<td>10</td>
<td>10</td>
<td>17</td>
<td>.00010</td>
<td>.00010</td>
<td>.00005</td>
<td>.55</td>
<td>.55</td>
<td>.55</td>
<td>12</td>
<td>5.30</td>
<td>1.44</td>
</tr>
<tr>
<td></td>
<td>660</td>
<td>250</td>
<td>10</td>
<td>12</td>
<td>25</td>
<td>.00055</td>
<td>.00025</td>
<td>.00010</td>
<td>.51</td>
<td>.51</td>
<td>.51</td>
<td>18</td>
<td>5.45</td>
<td>1.76</td>
</tr>
<tr>
<td></td>
<td>740</td>
<td>260</td>
<td>10</td>
<td>12</td>
<td>25</td>
<td>.00035</td>
<td>.00020</td>
<td>.00010</td>
<td>.51</td>
<td>.51</td>
<td>.51</td>
<td>15</td>
<td>5.02</td>
<td>2.11</td>
</tr>
<tr>
<td>C</td>
<td>720</td>
<td>260</td>
<td>10</td>
<td>11</td>
<td>13</td>
<td>.00025</td>
<td>.00010</td>
<td>.00010</td>
<td>.51</td>
<td>.55</td>
<td>.51</td>
<td>13</td>
<td>5.08</td>
<td>1.20</td>
</tr>
<tr>
<td></td>
<td>740</td>
<td>260</td>
<td>10</td>
<td>7</td>
<td>11</td>
<td>.00020</td>
<td>.00010</td>
<td>.00010</td>
<td>.52</td>
<td>.55</td>
<td>.51</td>
<td>14</td>
<td>5.08</td>
<td>1.33</td>
</tr>
<tr>
<td></td>
<td>740</td>
<td>260</td>
<td>10</td>
<td>12</td>
<td>18</td>
<td>.00020</td>
<td>.00010</td>
<td>.00010</td>
<td>.51</td>
<td>.55</td>
<td>.51</td>
<td>12</td>
<td>5.17</td>
<td>1.22</td>
</tr>
<tr>
<td></td>
<td>740</td>
<td>260</td>
<td>10</td>
<td>10</td>
<td>15</td>
<td>.00020</td>
<td>.00010</td>
<td>.00010</td>
<td>.51</td>
<td>.55</td>
<td>.51</td>
<td>12</td>
<td>5.22</td>
<td>1.30</td>
</tr>
<tr>
<td></td>
<td>680</td>
<td>250</td>
<td>10</td>
<td>9</td>
<td>16</td>
<td>.00025</td>
<td>.00010</td>
<td>.00010</td>
<td>.51</td>
<td>.55</td>
<td>.51</td>
<td>15</td>
<td>5.44</td>
<td>1.59</td>
</tr>
</tbody>
</table>

Information of the plausible range of each input parameter was retrieved from literature and computation for the need of modeling (details in section 2.5). Times \( t \) and \( t₀ \) are expressed in thousands of generations (assuming 20 years per generation). Population sizes \( N₀, N₁ \) and \( N₂ \) are expressed in thousands of individuals. \( D_{HC} \) represents the calculated average allele length difference on Human loci that are typed in Chimpanzee and \( D_{CH} \) in the reciprocal manner. Panel A shows best fits from an exploratory parameter search given a broad range of mutation rates (from \( 10^{-5} \) to \( 10^{-3} \)), while parameters \( b₀, b₁, b₂ \) and \( x \) are set as default values (\( b₀ = b₁ = b₂ = .55, x = 12 \)). The mutation rates in the top two best fits are below the generally accepted range. Although the other three fits yield acceptable mutation rate estimates, the parameter combinations result in very high probabilities of finding polymorphic loci, \( P(L \geq x) > .25 \). In Panel B, \( P(L \geq x) \leq .25 \) is assumed to ensure that the probability of choosing polymorphic loci is relatively small. Parameters \( b₀, b₁, b₂ \) are set equal and range from .51 to .55. \( x \) ranges from 12 to 18. The best fits are obtained when \( ν₂ \) is equal to the minimum possible value (\( 5 \times 10^{-5} \)), while fits become slightly worse if \( ν₂ \) is increased (\( 10^{-4} \)). In Panel C, when \( P(L \geq x) \leq .25 \) is still required while \( b₀, b₁, b₂ \) are allowed to vary independently, the parameter search tends to favor \( b₁ > b₂ \) and small \( x \) (< 15) to yield best fits. In both panels B and C with \( t, t₀, N, b \) and \( x \) assuming ranges of possible values when \( P(L \geq x) \leq .25 \), \( ν₀ \) is always greater than \( ν₁ \) and \( ν₂ \); \( ν₁ \) is greater than or equal to \( ν₂ \); \( ν₁ b₁ \) is always greater than \( ν₂ b₂ \).
2.4.3 Comparisons of empirical statistics derived from human and chimpanzee microsatellite data

We apply our model to analyze the well-known data set published by Cooper et al., (1998) [25]. These authors examined 40 human microsatellite markers and their homologues in a panel of non-human primates, and showed that human loci tend to be longer. Such trend was also confirmed by several other studies. Taken at face value, this data indicated that, since their most recent common ancestor, more microsatellite expansion mutations have occurred in the lineage leading to humans compared with the lineage leading to chimpanzees. Based on this, they suggested that this provided evidence that microsatellites tended to expand with time and were doing so more rapidly in humans. However, an alternative explanation, which attributes the difference to the influence of ascertainment bias, may also result in the observation of allele length difference. Therefore, Cooper et al. (1998) performed the necessary reciprocal experiment showing that human microsatellites tend to be longer than their chimpanzee homologues, regardless of the species from which the loci were cloned.

Dinucleotide (CA) repeat loci discovered and characterized in humans \((n = 22)\), were on average 5.18 repeat units longer than those in chimpanzees, while dinucleotide repeats discovered in chimpanzees \((n = 25)\) were on average 1.23 repeat units longer in humans. Table 1 lists best fits of three independent parameter searching results based on the genetic algorithm, with setup of \(X = 100\), \(Y = 1000\), probability of “crossover” = 0.6 and “mutation rate” = 0.02 in each search. Panel A shows best fits from an exploratory parameter search given a
broad range of mutation rates (from $10^{-5}$ to $10^{-3}$), while $b_0, b_1$ and $b_2$ and $x$ are set as default values ($b_0 = b_1 = b_2 = .55, x = 12$). The mutation rates in the top two best fits are below general accepted ranges, $\nu_2 = 2 \times 10^{-5} < 5 \times 10^{-5}$. Although the other three fits yield feasible mutation rate estimates, the parameter combinations result in very high probabilities of finding polymorphic loci, $P(L \geq x) > .25$, where $L$ denotes the incremental allele length of a sampled individual. See Remark below for derivation of $P(L \geq x)$. In Panel B, $P(L \geq x) \leq .25$ is assumed to ensure that the probability of choosing polymorphic loci is relatively small. $b_0, b_1, b_2$ are set to be equal and range from .51 to .55. $x$ ranges from 12 to 18. The best fits are obtained when $\nu_2$ is equal to the minimum possible value ($5 \times 10^{-5}$), while fits become slightly worse if $\nu_2$ is increased ($10^{-4}$). In Panel C, when $P(L \geq x) \leq .25$ is still required while $b_0, b_1, b_2$ are allowed to vary independently, the parameter search tends to favor $b_1 > b_2$ and small $x (< 15)$ to yield best fits.

For a range of evolutionary times, effective population sizes and mutation rates higher mutation rates, and rates of allele length expansions are always observed at Human microsatellite loci compared to those in Chimpanzee ($\nu_1 \geq \nu_2$ and $\nu_1b_1 > \nu_2b_2$), consistently with Cooper et al. (1998) [25] data.

2.4.3.1 Remark - derivation of $P(L \geq x)$

We denote random variable $L$ as the incremental allele length of a sampled individual, and show the computation of the expectation $E[L]$ and variance $V[L]$ given $t, t_0, \nu_0, \nu_1, b_0$ and $b_1$, where $t$ and $t_0$ are expressed in generation units.

Let $L_0, L_1$ be two random variables that denote the allele length increments in time
interval $t$ to $t_0$ (ancestral population 0) and $t_0$ to present (cognate population 1), respectively.

Let $X_i$ be the incremental change in the the $i$th generation of the ancestral population (from $t$ to $t_0$). We may obtain $P(X_i = 1) = \nu_0 b_0$, $P(X_i = -1) = \nu_0 (1 - b_0)$, $P(X_i = 0) = 1 - \nu_0$ and

$$E[X_i] = 2\nu_0 b_0 - \nu_0$$
$$V[X_i] = \nu_0 - \nu_0^2 (2b_0 - 1)^2$$

Based on the fact that the incremental change of allele length in any generation is independent of that in any other generation, we have

$$E[L_0] = (t - t_0) E[X_i] = (t - t_0) \nu_0 (2b_0 - 1)$$
$$V[L_0] = (t - t_0) V[X_i] = (t - t_0) \nu_0 [1 - \nu_0 (2b_0 - 1)^2]$$

Similarly we derive $E[L_1]$ and $V[L_1]$ as

$$E[L_1] = t_0 \nu_1 (2b_1 - 1)$$
$$V[L_1] = t_0 \nu_1 [1 - \nu_1 (2b_1 - 1)^2]$$

Therefore, we obtain

$$E[L] = E[L_0] + E[L_1] = (t - t_0) \nu_0 (2b_0 - 1) + t_0 \nu_1 (2b_1 - 1)$$
$$V[L] = V[L_0] + V[L_1]$$
$$= (t - t_0) \nu_0 [1 - \nu_0 (2b_0 - 1)^2] + t_0 \nu_1 [1 - \nu_1 (2b_1 - 1)^2]$$
Assuming that \( L \) is approximately Gaussian and given \( x \) as the threshold of allele size discovery, we calculate \( P(L \geq x) \) from the cdf of the Gaussian distribution. This also helps to determine a realistic range of estimates of \( t \) to control the probability of locus discovery.

**2.4.4 Influence of bottlenecks and expansions in human history**

While assuming a constant population size for Chimpanzee, we explore the influence of bottlenecks and expansions in human history on the observed difference in allele lengths (\( D \)). We extend the current modeling scheme and derive the analytical solution to compute \( D \) with Human cognate population size being arbitrarily varied from one generation to another.

Assume that the lineage of humans has been evolved following a multi-step demographic model, where there are \( L \) steps with human population size varied from step to step.

In the backward direction, we denote the present time in generation units as \( t_L = 0 \), \( t_{m-1} \) and \( t_m \) as the beginning and ending times of the \( m \)th step \((m = 1, 2, ..., L)\); \( N_m \) as the population size of the \( m \)th step. As already defined, \( t \) and \( t_0 \) are the age of the locus and the time when the two species split, respectively, and \( N_0 \) is the ancestral population size.

Chromosomes 1 and 1’ sampled at time 0 from population 1, have a common ancestor \( T \) units of time before present, either in population 1 at stage \( m \), if \( t_m \leq \)
Figure 2.4: Schematic representation of human demographic history with recent bottlenecks and expansions

Black line depicts human population, red line depicts ancestral and chimpanzee populations. \( t \): age of the locus (\( \sim 560,000 \) generations \( \sim 11.2 \) Myrs ago), \( t_0 \): species split (\( \sim 290,000 \) generations \( \sim 5.8 \) Myrs ago), \( t_1 \) : human migration out of Africa (\( \sim 10,000 \) generations \( \sim 200,000 \) yrs ago), \( t_2 \) : end of the last glaciation (\( \sim 600 \) generations \( \sim 12,000 \) yrs ago), \( t_3 \) : 0 A.D. (\( \sim 100 \) generations \( \sim 2,000 \) yrs ago), \( t_4 \) : beginning of industrialization (\( \sim 9 \) generations \( \sim 180 \) yrs ago).

\[
T \leq t_{m-1} \quad \text{(for } m = 1, 2, ..., L \text{)} \quad \text{or in the ancestral population } 0, \text{ if } T \geq t_0. \text{ Therefore,}
\]

\[
P(T > \tau) = \begin{cases} 
\exp \left[ - \sum_{k=m+1}^{L} \frac{(t_{k-1}-t_k)}{2N_k} \right] - \frac{(t_{m-1}-t_m)}{2N_m} & ; \quad t \leq t_{m-1}, \\
\exp \left[ - \sum_{k=1}^{L} \frac{(t_{k-1}-t_k)}{2N_k} \right] - \frac{(t_0-t_1)}{2N_0} & ; \quad \tau > t_0.
\end{cases}
\tag{2.16}
\]

for \( m = 1, 2, ..., L \). See Remark below for derivation of Equ. 2.15 in the extended model.

Taking a set of model parameters that fit the data from Table 2.1 \( t = 620,000, t_0 = 250,000, N_0 = N_1 = 10,000, N_2 = 17,000, \nu_0 = .0001, \nu_1 = .0001, \nu_2 = .00005, b_0 = b_1 = b_2 = 0.55, x = 12 \) we obtain \( D(H - C) = 5.30 \) and \( D(C - H) = 1.44 \) in the modeling scheme assuming fixed human population size.

Figure 2.4 depicts a schematic representation of major bottlenecks and expansions in the recent human history. The locus was born in the ancestral population, \( t \) generations ago. From \( t_0 \) when the two species split, effective population sizes for Human and Chimpanzee were equal to \( N_1 \) and \( N_2 \) (e.g. 5,000 and 20,000,
Burgess and Yang, 2008 [16]), respectively. At $t_1$ (∼ 200,000 yrs ago) when humans evolved to migrate out of Africa, there occurred a bottleneck event caused by fact that a subpopulation of migrants was sampled from a larger African population. Our stratified demographic model assumes that the decreased population size due to that bottleneck was constant until the end of the latest glaciation, $t_2$ (∼ 12,000 yrs ago). More precisely, it grew until the beginning of the last glaciation (∼ 50,000 yrs ago, Bond and Lotti, 1995 [13]) and then dropped, but the influence of this detail is minor. After that, human population underwent a series of expansions, with its effective size being ∼ $10^5$ from the end of last glaciation ($t_2$) to A.D. 0 ($t_3$ ∼ 2,000 yrs ago), ∼ $10^6$ from year 0 CE ($t_3$) to the emergence of industrialization ($t_4$ ∼ 180 yrs ago) and ∼ $10^8$ from $t_4$ to present time (current generation). Adapting the human demography with varying population sizes, as described above, in the extended model, we have calculated $D(H - C) = 5.42$ and $D(C - H) = 1.44$, compared to 5.30 and 1.44 obtained from the original model with fixed human population size. Using another set of model parameters from Table 2.1, $t = 720,000, t_0 = 260,000, N_0 = 10,000, N_1 = 11,000, N_2 = 13,000, \nu_0 = .00025, \nu_1 = .0001, \nu_2 = .0001, b_0 = b_2 = 0.51, b_1 = 0.55, x = 13$ results in $D(H - C) = 5.17$ and $D(C - H) = 1.20$ obtained from the extended model, compared with 5.08 and 1.20 obtained from the original model.

We conclude that for the range of parameters we considered, population bottlenecks and expansions in the recent human history have little impact on the modeled difference of allele sizes based on the settings of model parameters used in Table 2.1 to fit the data. Finally, the mutation rate estimate in the ancestral population
is consistently greater than that in Chimpanzee and the mutation rate estimate in Human is higher than or equal to that in Chimpanzee.

2.4.4.1 Remark - derivation of mean allele size difference, $D$, in the extended model

From the Eq. 2.16 derived above, we obtain

$$f_T(\tau) = \begin{cases} 
(\frac{1}{2N_m})e^{\exp[-\sum_{k=m+1}^L (\frac{t_{k-1-\frac{t_k}{2N_k}})}{2N_k}) - \frac{\tau-t_m}{2N_m}]}; & t < \tau \leq t_{m-1}, \\
\exp[-\sum_{k=1}^L (\frac{t_{k-1-\frac{t_k}{2N_k}})}{2N_k}) - \frac{\tau-t_0}{2N_0}]}; & \tau > t_0.
\end{cases}$$

If $\tau \leq t_0$ and $t_m < \tau \leq t_{m-1}$ ($m = 1, 2, ..., L$), we obtain

$$E[Y'_1; Y_1 = i|T = \tau] = e^{-\nu_0(t-t_0)-\nu_1t_0} \{[\nu_0(t-t_0)b_0 + \nu_1(t_0-\tau)b_1]_{\beta_i} - \nu_1(t-t_0)b_1 - [\nu_0(t-t_0)d_0 + \nu_1(t_0-\tau)d_1]_{\beta_{i+1}}\}$$

$$E[Y'_1; Y_1 = i] = \sum_{m=1}^L \int_{t_m}^{t_{m-1}} E[Y'_1; Y_1 = i|T = \tau](\frac{1}{2N_m})e^{\exp[-\sum_{k=m+1}^L (\frac{t_{k-1-\frac{t_k}{2N_k}})}{2N_k}) - \frac{\tau-t_m}{2N_m}]d\tau$$

where $\beta_i$ is the same as that defined and used in the derivation of Eqs. (2.10) and (2.11).

If $t_0 < \tau \leq t$, we obtain

$$E[Y'_1; Y_1 = i|T = \tau] = e^{-\nu_0(t-t_0)-\nu_1t_0} \{[\nu_0(t-\tau)b_0]_{\beta_i} - \nu_0(t-\tau)d_0\}$$

$$E[Y'_1; Y_1 = i] = \int_{t_0}^t E[Y'_1; Y_1 = i|T = \tau](\frac{1}{2N_0})e^{\exp[-\sum_{k=1}^L (\frac{t_{k-1-\frac{t_k}{2N_k}})}{2N_k}) - \frac{\tau-t_0}{2N_0}]d\tau$$

57
If $\tau > t$, we obtain

$$E[Y'_1; Y_1 = i|T = \tau] = e^{-\nu_0(t-t_0)} - \nu_1 t_0 [\nu_0(t - t_0)(b_0 - d_0) + \nu_1 t_0 (b_1 - d_1)] \beta_i$$

which is not a function of $\tau$.

Therefore

$$E[Y'_1; Y_1 = i] = \int_t^\infty E[Y'_1; Y_1 = i|T = \tau] f_T(\tau) d\tau = E[Y'_1; Y_1 = i|T = \tau] P(T > t)$$

$$= e^{-\sum_{k=1}^L (\frac{t_{k+1} - t_k}{2N_k}) - \nu_0 t_0 - \nu_1 t_0 [\nu_0(t - t_0)(b_0 - d_0) + \nu_1 t_0 (b_1 - d_1)] \beta_i}$$

With the analytical derivation shown here we are able to compute the allele length difference $D$ with Human population size arbitrarily varied from generation to generation.

### 2.5 Discussion and conclusions

Computations presented in this chapter demonstrate that the scaled forward simulations using simuPOP closely match the analytical solution of the evolutionary model used. We note that mathematical derivation of equation (2.15) depends on simplicity of the assumed microsatellite discovery criterion $Y_1 \geq x$. If this criterion is replaced by a condition on heterozygosity or variance, the theoretical derivations become very difficult. On the other hand, it is easy to use any other microsatellite discovery criterion in simuPOP simulations.

Data of Cooper et al. (1998) analyzed that when the human-derived dinucleotide repeat loci are typed in chimpanzee, they showed a trend towards smaller
mean allele sizes in the Chimpanzee as compared to that in human populations. These and other data also suggest that the same holds for other measures of within-population variation (i.e., the chimpanzees showing lower heterozygosity and allele size variance, compared to humans, Vowles and Amos (2006) [115]). The theoretical model shows that these observations are in agreement with the presence of ascertainment bias, caused by a selective choice of human loci. In the reciprocal experiment, the chimpanzee-derived dinucleotides, typed in human populations, also show a trend towards smaller mean allele sizes in the chimpanzee as compared to that in human populations.

We adapted a genetic algorithm (Mitchell, 1996) [82] to perform an extensive parameter space search by specifying a number of values of each of the key modeling parameters \((t, N, \nu, b \text{ and } x)\), see Table 2.1 for details) which are variable within plausible ranges. Patterson et al. (2006) [88] reviewed the estimated times of divergence of the two species \((t_0)\) and determined that divergence occurred approximately between 250,000 and 350,000 generations ago. This corresponds to approximately \(\sim 5\) to 7 million years by assuming 20 years per generation. For the purpose of modeling, the time when a particular locus was born \((t)\) is computed to be varying around 450,000 to 750,000 generations to ensure the threshold of allele size being large enough so that the polymorphic locus occurs only rarely \((\leq 25\% \text{ of loci, see section 2.4.3.1 for details})\). Using both likelihood and Bayes methods, Yang (2002) [124], estimated that the ancestral \((N_0)\) and Chimpanzee \((N_2)\) effective population sizes were ranging from 10,000 to 20,000 individuals, and the Human effective population size was ranging from 3,000 to 12,000 individuals (Burgess and
Yang, 2008 [16]). Chen and Li (2001) [21] suggested a much larger effective population size, 50,000 ∼ 90,000, of the common ancestor of Human and Chimpanzee.

We assign multiple numbers within these ranges as possible values of $N_0, N_1$ and $N_2$. Additionally, given that microsatellite loci mutation rate in any population is greater than $10^{-4}$, as analyzed by Ellegren (2000) [38], $\nu_0, \nu_1, \nu_2$ are assumed in a wide range starting from $5 \times 10^{-5}$. The presence of mutational bias (Sainudiin et al., 2004; Wu and Drummond, 2011) [104, 121] has been represented by $b_0, b_1, b_2$, which are ranging from .51 to .55 and denote unequal probabilities of expansion and contraction for mutational events. In this model, we assume such bias to be constant within a population, where $b$ is constant for all alleles in one population and can be a different constant in another population. As demonstrated in Table 2.1 for a range of effective population sizes and evolutionary times, the estimated Human mutation rates are always higher than or equal to those in Chimpanzee and the mutation rate estimates in the ancestral population are always greater than those in either Human or Chimpanzee.

These observations imply that ascertainment bias is a significant factor in interpreting inter-population genetic variation at microsatellite loci, when the loci are selectively chosen for polymorphism in one of the populations compared. Ascertainment bias effect is confounded by other differences in evolutionary dynamics between the cognate and non-cognate populations, particularly by inter-population differences of rates of mutations at the locus. As shown in Figure 2.3A, increased mutation rate in the non-cognate population reduces the effect of the ascertainment bias, while increased mutation rate in the cognate population amplifies the
effect of the bias (Figure 2.3b). On the other hand, the primary cause of ascertainment bias is a tighter correlation of allele sizes within the cognate population. Thus, intuitively it is clear that population size differences between cognate and non-cognate populations also may reduce or amplify the ascertainment bias. If the cognate population is of larger size or is growing more rapidly than the non-cognate one, a reduced bias is expected.

The differences of patterns of biases seen at the dinucleotide loci discovered in Human versus Chimpanzee, can be explained in the terms of the theoretical predictions, if the mutation rate is higher for humans. The observed pattern that ascertainment bias is of a lower magnitude for the chimpanzee-specific loci is also consistent with effective population size in chimpanzees being smaller than that of human populations. In this sense, our observations and theoretical predictions are consistent with the assertion of Rubinsztein et al. (1995) [103], although expansion bias of mutations is not necessary to explain the observed differences in humans and chimpanzees.

As mentioned, when describing the model, the time and mutation rates (as well as effectively the population sizes) are scaled to the unit equal to the the human generation length. This is convenient, and the numbers can be easily rescaled to accommodate different evolutionary parameters in different species. Our theory and data can also be used to explain the apparently discordant conclusions reached by other investigators examining this issue. For example, Ellegren et al. (1995) [39] observed smaller allele sizes in non-cognate species compared with cognates of birds, which could be predominantly due to ascertainment bias alone. Crawford et
al. (1998) [26], in contrast, found longer median allele sizes in sheep compared with cattle, regardless of the origin of the microsatellites. This may be the case where the ascertainment bias effect is counteracted or even reversed due to mutation rate and (or) effective population size differences in sheep and cattle.

There had been discussions with regard to the dependence of inter-population allele size differences on the absolute repeat lengths of alleles (Ellegren et al., 1995; Amos and Rubinsztein, 1996) [39, 6]. For microsatellites, there is a general tendency for increased level of polymorphism at loci harboring larger alleles (Weber, 1990) [116]. Our theory shows that loci exhibiting higher degrees of polymorphism will be subject to lesser bias of ascertainment (due to lower correlation of allele sizes in the cognate population). Hence, appropriate adjustment of inter-locus differences of polymorphism as well as allele sizes should be made in addressing the importance of ascertainment bias.

Vowles and Amos (2006) [115] is an important contribution to the literature on ascertainment bias. Among other, these authors observed that long repeats tend to be interrupted, which contributes an additional bias. They also proposed that the difference $D$ be explained if microsatellites evolve at different rates, with longer microsatellites evolving faster; this latter effect having some statistical rationale.

In this chapter, we offer an explanation which does not rely on interruption nor acceleration, but only on sampling, and demographic and population-genetic effects, under constant though species-dependent mutation rates. However, there is at least some concordance; we find that human microsatellites, which are on the average longer, also have higher mutation rates, which might be a hint that both approaches
detect the same or similar effect.

In summary, we conclude that ascertainment bias is an important consideration for interpretation of inter-population differences of genetic variation at microsatellite loci, but this bias can be reduced or even reversed when the past demographic histories of cognate and non-cognate populations are different. In addition, mutation rate differences in populations either due to their reproductive behavior or differences of cell division during oogenesis and spermatogenesis can also influence or mimic ascertainment bias.
Chapter 3

Simulation framework to generate and analyze sequence-based data for rare variant association studies

3.1 Motivation and overview

Currently, there is great interest in detecting complex trait rare variant associations using next-generation sequence data. On a monthly basis, new rare variant association methods are published. It is difficult to evaluate these methods because there is no standard to generate data and often comparisons are biased. In order to fairly compare rare variant association methods, it is necessary to generate data using realistic population demographic and phenotypic models. We build an interactive program, SimRare, which integrates generation of rare variant genotype/phenotype data and evaluation of association methods using a unified platform. Variant data can be generated for gene regions using forward-time simulation that incorporates realistic population demographic and evolutionary scenarios. Phenotype data can be obtained for both case-control and quantitative traits. SimRare is designed to have a user-friendly interface that allows for easy entry of genetic and phenotypic parameters. Novel rare variant association methods implemented in R program-
ming language can also be imported into SimRare, to evaluate their performance and compare results, e.g. power and Type I error, with other currently available methods both numerically and graphically. Using SimRare we validate existing association methods with data generated under various scenarios of demographic history and disease etiology. We demonstrate that the power of each method depends on the underlying model and differences in power between methods are usually modest.

3.2 Introduction

Next generation sequencing technologies, e.g. Illumina HiSeq, ABI 454, Roche SOLiD have made it possible to cost-effectively identify rare variants through the generation of exome and whole genome sequence data. It has been demonstrated that for complex traits testing individual rare variants is grossly underpowered (Gorlov et al., 2008) [51]. Therefore a large number of rare variant association methods, which aggregate variants across a region, which is usually a gene, have been developed.

However, comparing the power of these methods to detect associations by reviewing the literature is difficult, because there is no regularity in data generation and sometimes to make a particular method appear more powerful than others data are generated in an unrealistic manner. In order to fairly compare rare variant association methods it is necessary to generate data using realistic population demographic and phenotypic models and compare methods via data generated using the same parameters. Although variant data can be obtained for example from
the 1000 Genomes Project (The 1000 Genomes Project Consortium, 2010) [1] and
NHLBI-Exome Sequencing Project (Tennessee et al., 2012) [112], due to limited
samples sizes, it is still not possible to generate variant data which is reflective of
the true distribution of rare variants (e.g. singletons, doubletons). It is therefore
crucial to be able to incorporate evolutionary history to generate large samples of
variant data.

The SimRare program was developed to evaluate type I and II error of rare variant
complex trait association methods by generating both variant and phenotype data
using realistic models. As a unified simulation platform with user-friendly interface
for rare variant studies, SimRare provides an unbiased and easy manner to evaluate
association methods, including novel methods. It consists of three modules, vari-
ant data simulator, genotype/phenotype generator and association method evalua-
tor. SimRare generates variant data for gene regions using forward-time simulation
which incorporates realistic population demographic and evolutionary scenarios.
For phenotype data it is capable of generating both case-control and quantitative
traits. The phenotypic effects of variants can be detrimental, protective or non-
causal. SimRare which has a graphical interface is written in Python/C++ and
can incorporate R scripts for the evaluation of novel and existing methods.

3.3 Description

3.3.1 Generation of Variant Data

SimRare can generate variant data using several existing population genetics and
demographic models (e.g. Boyko et al., 2008; Kryukov et al., 2009) [15, 72] or
parameters for other models can be specified. Variant data is generated using forward time simulation (Peng and Kimmel, 2005; Peng and Liu, 2010) [91] [92]. This implementation is superior to other methods (Liu et al., 2008) [78] in that it incorporates a realistic evolutionary model, which includes demography, variable mutation rates, recombination, multi-locus fitness effect and locus-specific selection models including purifying selection. Using these population demographic models it is possible to generate variant data, for different populations, e.g. European and African.

3.3.2 Generation of Phenotype Data

To evaluate type I it is only necessary to generate data under the null hypothesis, i.e. trait and variants are not associated; therefore it suffices to generate variant data for a region, unconditional on the phenotype. To evaluate Type II error (power) it is necessary to generate trait data conditional on the variant data within a region. Either conditional or unconditional on the generated variant data SimRare can generate phenotype data for both case-control and quantitative traits using a wide range of models. SimRare can generate phenotype data for both case-control and quantitative traits for a wide range of genetic etiologic assumptions, including penetrance-prevalence model, liability threshold model, population attributable risk and linear model. The underlying genetic mode of inheritance can be specified. For quantitative traits a random or an extreme trait sample can be generated. When an extreme quantitative trait sample is generated the ranges can be specified, e.g. upper 90% and lower 10%. All sample sizes can be user specified. Causal variants within a “gene” region can either have an effect which is unidirectional e.g. all
variants increase disease risk or are bidirectional, e.g. can either decrease or increase quantitative trait values. For the causal variants the magnitude of the effect can either be fixed, i.e. all causal variants have the same effect size or variable, effect size is determined by frequency or selection coefficients.

3.3.3 Analysis of Data and Evaluation of Type I and II error

It is possible to analyze the generated data using a wide range of complex trait rare variant association methods, e.g Combined Multivariate and Collapsing method (Li and Leal, 2008) [76], Weighted Sum Statistic (Madsen and Browning, 2009) [80], Variable Threshold (Price et al., 2010) [95]; Kernel-Based Adaptive Cluster (Liu and Leal, 2010) [77]. Additionally any newly developed rare variant association method written in R can be evaluated using SimRare. For case-control studies and randomly ascertained quantitative trait studies, qualitative and quantitative analysis is performed respectively. While for an extreme quantitative trait sample, the data can be dichotomized and qualitative analysis performed or the qualitative trait values can be analyzed. For the evaluation of type I and II error the user determines the number of replicates which should be generated. For many of the rare variant association methods significance cannot be evaluated analytically, therefore for empirical evaluation the number of permutation and significance level can be specified when evaluating type II error.

3.3.4 Program overview

SimRare, a stand-alone executable software, is compiled for Windows, Mac and Linux operating systems. It only requires installation of R if novel methods are
desired to be evaluated. SimRare consists of three major modules: simulation of sequence-based genotype data, generation of phenotype data and evaluation of association methods. SimRare has a user-friendly interface. Output from SimRare can either take the form of tables or graphs, e.g. QQ plots for type I errors and bar graphs to display power. Additionally simulated data can be written into external files in standard linkage format. A benchmark test was performed on a Ubuntu Linux 64-bit machine with Intel 3.2 GHz CPU and 8GB RAM. First 100 haplotype pools were generated using a population demographic model (Kryukov et al., 2009) [72]; the time to generate each pool is 1.8 minutes. The generated haplotype pools can be used for all subsequent power analyses. One thousand replicates were generated each with 1,000 case and 1,000 controls and power was evaluated using the CMC method (Li and Leal, 2008) [76]. The total computational time to perform the power analysis was 18.4 minutes.

3.3.4.1 Where to apply

**Benchmark** - fairly and easily evaluate rare variant association methods to assess their power and robustness.

**Study design** - aid parameter analysis of underlying disease model to test a wide variety of scenarios or to achieve adequate power.

**Method development** - boost statistical geneticists in design of novel association methods without having to develop special software to generate data in order to evaluate their methods.
3.4 Program Guide

3.4.1 Module of variant data generation

This module repeatedly evolves a population of gene-based sequences using forward-time simulation with mutation, natural selection and demography. At the end of each repeated simulation, variant sites information such as minor allele frequencies, selection coefficients, position information will be saved. Overall, running this module will create a variant haplotype data pool for a gene region based on any user-defined evolutionary scenario. This simulation module is featured by its capability of incorporating multi-stage population expansion/bottleneck model, handling multi-locus selection model of fitness and random or locus-specific selection coefficient to novel variant, and creating reusable gene-based variant data pool.

This module is launched by “Need generate rare variant data pool? –> Click Here”, which is shown on the SimRare main window. A snapshot of the interface is shown in Figure 3.1.

3.4.1.1 Basics

Users may choose to load previously saved configuration file or use parameter defaults to avoid inputting parameters manually. For loading a configuration file, check “Use Saved Input Parameters?” and click “Load File”. To use default input parameters, select “Use Default Input Parameters?”.

Note. First time user always need to input parameters or choose to use defaults.

The configuration file can be generated by clicking “Save Input Parameters” and saved for future use. Manual input of parameters begins with inputting
This module is implemented in simuPOP (Peng and Kimmel, 2005) and adapts features from srv (Peng and Liu, 2011), a program implemented in simuPOP to simulate sequences of the human genome with rare variants.

basic information, such as gene length, name of output files and number of replicates.

- “Gene Length Mode” – use “Fixed” to specify a number as fixed gene length for each replicate; select “Random” to choose a random number between a specified range for each replicate as its gene length

- “Gene Length” – if “Fixed”, specify a fixed gene length for all replicates, e.g. 2500, otherwise, input the range of gene length, e.g. 2500 | 5000

- “Output Files Name (prefix)” – input a file name as the prefix of output files, e.g. simVarPool

- “Number of Replicates” – input a positive integer, which is equivalent to the size of the data pool, e.g. 100
3.4.1.2 Demography

For a $N$-stage demographic model, effective population sizes at the beginning of each stage and at the end of the last stage are required along with specification of number of generations at each stage.

- “Effective Population Sizes” – a sequence of $(N+1)$ positives integers separated by either space or comma or semicolon, where the first $N$ numbers specify the population sizes at the beginning of each stage and the last number indicates the population size at the end of the last stage. For any two adjacent numbers, $i$&$j$: if $i = j$, the population size remains constant through the corresponded stage, such as a burn-in stage, etc.; if $i > j$, the population size is reduced to $j$ instantly by entering that stage to mimic bottleneck effect; if $i < j$, the population expands exponentially from $i$ to $j$ through that stage.

  – E.g. “5000, 5000, 800, 30000”, which simulates a three-stage demographic model where a population beginning with 5000 individuals, first undergoes a burn-in stage with constant population size 5000, then a bottleneck of 800 individuals, and after that expands exponentially to a size of 30000.

**Note.** If “Number of Generations Per Stage” has already been specified as a sequence of $M$ numbers, a sequence of $(M + 1)$ numbers is required as input of “Effective Population Sizes”.

- “Number of Generations Per Stage” – a sequence of $N$ positive integers separated by space or comma or semicolon, where the $i$th value represents the number of generations at the $i$th stage.
E.g. “300, 100, 200” specifies a three-stage demographic model where the population is evolved 300, 100 and 200 generations through the 1st, 2nd and 3rd stages, respectively.

**Note.** If “Effective Population Sizes” has already been specified as a sequence of \((N + 1)\) numbers, a sequence of \(N\) number is required for “Number of Generation Per Stage”.

### 3.4.1.3 Genetic forces

Users opt to determine mutation model, mutation rate, locus-specific selection model, selection coefficient distribution model and recombination.

- “Mutation Model” – select “infinite\_sites” for infinite-sites mutation model, where mutation can occur only at wild-type loci; select “finite\_sites” for finite-sites mutation model, where mutation can occur at any site.

- “Mutation Rate” – mutation rate per base pair.

- “Revert Fixed Sites” – if checked, the program reverts fixed sites to wild type alleles during the evolutionary simulation, otherwise fixed sites will remain as variant sites.

- “Selection Coefficient Distribution” – choose one of commonly used models as the distribution to assign selection coefficient to any new mutant in order to determine its site-specific fitness of genotypes \(AA, Aa/aA, aa\), where \(A\) is wild type allele.

- If “Constant” (Williamson *et al.*, 2005) [119] is selected it sets all mutant
sites a constant selection coefficient 0.01, whereas other options use the estimated selection parameters from corresponded demographic models, such as “Boyko_2008_European, Boyko_2008_African, Eyre-Walker_2006 and Kyrukov_2009” [15, 44, 72].

**Note.** In single-locus selection model, fitness of genotypes $AA, Aa/aA$ and $aa$ are given by $1$, $1 - hs$ and $1 - s$, respectively, where $s$ is the selection coefficient and $h$ is the dominance coefficient (default 0.5 for additivity). Also, $s > 0$ results in site being negatively selected, $s = 0$ in neutral and $s < 0$ in positively selected.

- “Multi-locus Selection Model” – choose from additive, multiplicative and exponential to determine the mode to calculate an individual’s fitness over all of its variant sites.

- “Customized Selection Coefficient” – (optional) specify any customized selection coefficient distribution model other than selecting a commonly used one from “Selection Coefficient Distribution”. The input slot is enabled by checking “Need Customize Selection Coefficient?”. There are three modeling schemes available (constant, gamma distributed and mixed-gamma distributed). For constant, it requires a list of two values, $s, h$, constant selection coefficient and dominance coefficient (e.g. “-0.001, 0” specifies a recessive model with fixed positive selection). A list of three input values is required for gamma distribution, $k, d, h$, where $k, d$ are shape and scale parameters of the distribution and $h$ is the dominance coefficient (e.g. “0.206, 0.292, 0.5” specifies a gamma distribution with parameters 0.206, 0.292 and with selection coefficient $h = 0.5,$
which replicates “Boyko_2008_European” model [15] and has the same effect as choosing “Boyko_2008_European” from “Selection Coefficient Distribution”. For specifying mixed-gamma distribution model, an acceptable input is a list of either 5 or 7 parameters which are \( p, s, k, d, h \) or \( p, s, k, d, h, l, u \), where \( p \) is the probability of having the selection coefficient equal to \( s \); \( k, d \) are gamma distribution parameters; \( h \) is the dominance coefficient and \( l, u \) (optional) are lower and upper boundaries of the selection coefficient.

- “Recombination Rate” – (optional) use “Need Recombination ?” to enable the input slot for recombination rate per base pair.

**Note.** If the specified recombination rate times gene length being greater than 0.5, a rate of 0.5 divided by gene length will be used.

### 3.4.1.4 Save and run

- “Screen Output Mode” – choose one from quiet (no screen output), minimum (minimum output of simulation progress and time spent for each replicate) or regular (regular screen output of statistics, simulation progress and time spent, etc.)

- “Detailed Screen Output Interval per Stage” – enabled only if “regular” is selected for “Screen Output Mode”; a single number or a list of \( n \) numbers is required (\( n \) equals to number of demographic stages) as intervals of number of generations at which statistic information will be calculated and output. If unspecified, summary statistics from the beginning to the end of every generation at each stage will be shown on the screen.
• “Save Genotype for Replicates” – (optional) output genotype information. This option is disabled by default because this format is not efficient in storing rare variants. If selected and specified by a positive number $n$, genotype in standard linkage format for the first $n$ replicates will be output to $n$ files, i.e. 1: genotype of the first replicate will be saved to file; 3: genotype of the first three replicates will be output to three files; -1: genotype information of all replicates will be saved.

• “Save Statistics for Replicates” – (optional) output statistics (same input as “Save Genotype for Replicates” requires).

• “Set Path to Output Files” – (optional) change output file path to a desired directory. The default is current directory where SimRare is launched.

• “Save Input Parameters” – (optional) save user inputs into an external parameter file (*.par).

3.4.1.5 Output files

The standard output of this module includes four files which are *.maf, *.sel, *.pos and *.len. Number of lines in each of them equals to the “Number of Replicates” (assume number of replicates is $n$). For $i = 1, 2, ..., n$, the $i$th line in len file is the gene length of the $i$th replicate, the $i$th line in pos file contains $n_i$ values of position information of mutant sites, where $n_i$ is equal to the number of total mutant sites for the $i$th replicate, and the $i$th lines in both maf and sel files have $n_i$ values that are minor allele frequencies and site-specific selection coefficients on corresponded $n_i$ mutant sites for the $i$th replicate.
Additionally, while running any replicate during simulation users have the option to save genotype information of all individuals into file in standard linkage (ped) format with file name as “fileName_rep_i.ped”, where “fileName” is specified by “Output Files Name (prefix)”. It is essential to obtain maf, sel and pos files before launching any other SimRare module.

3.4.2 Module of genotype and phenotype generation

Based on the generated genotype from the simulated pool of variant data, this module can generate phenotype data for both case-control and quantitative traits. Users have a wide range of choice among available underlying study design models, such as case-control odds ratio and prevalence (Risch, 1990) [99], population attributable risk (Madsen and Browning, 2009) [80], quantitative traits (Kryukov et al., 2009) [72], QTL extreme traits sampling (Huang and Lin, 2007) [58] and Mendelian traits (Ott, 1999) [86] simulation models.

A snapshot of the interface of this module is shown in Figure 3.2

3.4.2.1 Initialization

In order to simulate variant sites based genotype and phenotype data we need to first import previously generated maf (minor allele frequencies), sel (site-specific selection coefficients) and pos (variants positions information) files.

- “Use Saved Parameters” – (optional) open a saved par (input parameters) file and import previously specified user inputs (see section 3.4.1.4 for how to create a par file).

- “Choose a study design model” – select a study design model from “How do
Figure 3.2: Snapshot of SimRare interface - module of generating genotype & phenotype data
you want to establish genotype-phenotype associations?” There are six available study designs currently to choose from: case-control prevalence model of the population, case-control prevalence model of a sample, case-control population attributable risk model of a sample, quantitative-traits (QT) simulation, quantitative-trait-locus (QTL) extreme sampling, and Mendelian traits simulation. Once the model has been selected required input slots for model-dependent parameters will be enabled accordingly.

- “Use Default Input Parameters” – (optional) automatically fill in input parameters with default values.

### 3.4.2.2 Model parameters

Only those that are enabled are required input parameters for the selected study design model. Please refer accordingly.

- “Proportion of Detrimental RVs” – proportion of functional detrimental/deleterious variants.

- “Proportion of Protective RVs” – proportion of functional protective/beneficial variants.

**Note.** Each variant site is determined as synonymous, detrimental or protective by its corresponding selection coefficient.

- “# Cases” – number of cases.

- “# Controls” – number of controls.

- “# Unphenotyped” – number of unphenotyped cohort controls.
• “# Samples” – total number of samples.

• “Fixed Effect Model/Variable Effect Model” – choose “Fixed Effect Model” to set odds ratio per causal variant to be a constant, or “Variable Effect Model” to set odds ratio per casual variant to be determined by its minor allele frequency.

• “Prevalence” – disease prevalence (baseline penetrance of a gene).

• “Odds Ratio for Common Mutations” – odds ratio for common variants (1.0 for neutral, > 1.0 for deleterious and < 1.0 for protective).

• “Odds Ratio for Protective Mutations” – specify odds ratio (< 1.0) for protective variants (a constant for “Fixed Effect Model” or a value range for “Variable Effect Model”).

• “Odds Ratio for Detrimental Mutations” – specify odds ratio (≥ 1.0) for detrimental variants (a constant for “Fixed Effect Model” or a value range for “Variable Effect Model”).

• “Mode of inheritance” – choose the mode of inheritance.

• “Attributable Risk for Detrimental Mutations” – total population attributable risk for deleterious variants.

• ’Attributable Risk for Protective Mutations’ – total population attributable risk for protective variants.

• “Constant Parameters?” – (optional) set locus-specific population attributable risk inversely proportional to its minor allele frequency rather than uniformly distributed.
• “QT Coefficient for Common Variants” – mean value shift for common variants.

• “QT Coefficient for Causal Variants” – specify mean value shift for causal variants (a constant for “Fixed Effect Model” or a value range for “Variable Effect Model”)

• “QTL Cutoffs” – lower and upper percentile cutoffs for quantitative traits in extreme QT sampling.

• “Mark Case-Control?” – (optional) convert extreme quantitative traits into case-control traits by recoding extreme quantitative traits using the binary coding.

• “Percentage of Causal RVs” – percentage of rare variants being causal in Mendelian traits simulation.

• “Proportion of Heterogeneous Cases” – proportion of cases that do not carry the disease allele at the gene region.

• “Allelic Heterogeneity?” – (optional) if left unselected there is no allelic heterogeneity for Mendelian traits, otherwise it will fix the causal variants of the Mendelian trait to the one that has the (proportion of heterogeneous cases)x100%-th smallest minor allele frequency.

• Study design model specificity

  – “Case-control prevalence model of the population” – need to specify proportions of functional RVs (detrimental and protective), # samples, prevalence, effects model, odds ratios (common, detrimental & protective variants) and mode of inheritance.
- “Case-control prevalence model of a sample” – need to specify proportions of functional RVs, # cases, # controls, # unphenotyped, effects model, prevalence, odds ratios and mode of inheritance.

- “Case-control population attributable risk model of a sample” – need to specify proportions of functional RVs, # cases, # controls, # unphenotyped, mode of inheritance, attributable risks (detrimental and protective mutants) and if locus-specific risk is constant.

- “Quantitative-traits (QT) simulation” – need to specify proportions of functional RVs, # samples, effects model, QT coefficients (common and causal variants).

- “QTL extreme traits sampling” – need to specify proportions of functional RVs, # unphenotyped, # samples (or # cases and # controls), effects model, QT coefficients, QTL cutoffs, if quantitative traits will be marked as case-control.

**Note.** Under this study design model, there are two sampling schemes to sample either a number of cases & controls or a number of cohort samples. If “# cases” and “# controls” are specified the simulator generates exact numbers of cases and controls as specified that have QT extreme traits, whereas if “# samples” is specified, it draws out all individuals that have trait-specific extreme values as cases or controls from the cohort samples (numbers of cases and controls are unknown before the simulation completes but definitely smaller than the specified number of cohort samples).
“Mendelian traits simulation” – need to specify # cases, # controls, mode of inheritance, % causal RVs, if allelic heterogeneity is invoked and proportion of heterogeneous cases.

3.4.2.3 Mimic genotyping

To mimic the actual genotyping process that involves missing sites the estimated proportions of different types of missingness can be specified within the area of “Mimic Genotyping Process”.

There are four types of missingness, “missing detrimental variants”, “missing protective variants”, “missing non-causal mutants” and “missing synonymous mutants”, respectively. Missing genotypes will be encoded as wildtypes by default.

- “Mark Missing RVs” – (optional) recode missing data from wildtype genotype (0) to -9 to indicate the missingness.

3.4.2.4 Save and run

- “Save Input Parameters?” – save user inputs to a parameter configuration (par) file. A saved par file can be imported by using “Use Saved Input Parameters”.

- “Is Syno Trimmed?” – (optional) remove synonymous variant sites.

- “Is CV Trimmed?” – (optional) remove common variant sites.

- “Is Ped Written?” – (optional) save a simulated genotype/phenotype data set in linkage format (ped file).

- “Print Genotypes?” – (optional) print out the generated genotype on screen.

- “Print Phenotypes?” – (optional) print out the generated phenotype on screen.
• “Print Minor Allele Frequencies?” – (optional) print out variants’ minor allele frequencies on screen.

### 3.4.3 Module of association tests

This module provides most commonly used complex trait rare variant association methods to choose from and to compare each other’s performance, such as 1) Combined Multivariate and Collapsing Method (CMC) (Li and Leal, 2008) [76]; 2) Weighted Sum Statistic (WSS) (Madsen and Browning, 2009) [80]; 3) Variable Threshold (VT) (Price et al., 2010) [95]; etc. It is also capable of incorporating any novel association method that is developed in R program and comparing it with existing ones. Users are required to specify testing thresholds, such as significance level and number of replicates to be generated. A large number of existing tests are permutation based because the significance of their association methods cannot be evaluated analytically. To apply those tests users also need to specify number of permutations and should expect a large load of computational burden.

A snapshot of this module is shown in Figure 3.3

### 3.4.3.1 Configuration

Click “Load configuration File” to choose the desired parameter (par) file that has been created by the module of genotype & phenotype simulation (see section 3.4.2.4).

• “Rare variant frequency bound” – set lower and upper bounds of observed sample minor allele frequency. Loci that have observed MAFs being out of the boundary will not be analyzed.
To activate this module, click “Need apply existing/novel association methods? –> Click Here” shown on the upper right corner of SimRare main interface (Fig. 3.2).

- “Specify number of replicates for power calculation” – input number of replicates to simulate.
- “Specify Significance level” – at which power will be evaluated (< 0.5).
- “Specify number of permutations” – input number of permutations to run for permutation based methods.

### 3.4.3.2 Existing methods

At least one of existing methods has to be selected or a novel method (written in R script) has to be imported.

- “CMC” – (Li and Leal, 2008) \[76\], Combined Multivariate and Collapsing, use “CMC-one” for one-sided test or use “CMC-QT” for test of quantitative traits.
• “MZ” – (Morris and Zeggini, 2010) [83], search of RV accumulations within the same functional unit, choose “MZ-one” for one-sided test or “MZ-QT” for test of quantitative traits.

• “WSS” – (Madsen and Browning, 2009) [80], Weighted Sum Statistic, choose “WSS-one” for one-sided test.

• “VT” – (Price et al., 2010) [95], Variable Threshold, for one-sided test select “VT-one”.

• “KBAC” – (Liu and Leal, 2010) [77], Kernel-Based Adaptive Cluster, choose “KBAC-one” for one-sided test.

Note. “CMC-QT”, “MZ-one”, “MZ”, “KBAC-one”, “KBAC”, “VT-one”, “VT” are permutation based tests which can cost a lot of computational time.

3.4.3.3 Novel method

Novel association method implemented in R programming language can be imported and evaluated with other existing methods. Choose “Load R file of a new method” and then click “Load File” to select the r file.

Open a terminal (or command prompt in Windows) and try if command “Rscript” works. If it is not recognized as a command, set R to system PATH first, and then follow the steps below.

1. Include a few lines of code on top of the R script to read the ped file (“pedFile”), which denotes the full path to the file that saves the simulated data, and to retrieve phenotypes and genotypes from it. For example, add these lines to the very beginning of your R script
# read ped file as a data matrix
simData <- as.matrix(read.table(pedFile, header=FALSE, sep=" "))

# first 5 columns are irrelevant to our need, phenotypes are stored in the 6th col
# genotypes are stored from the 7th to the last column
pheno = simData[, 6]
geno = simData[, -(1:6)]

Note. “pedFile” is a fixed name not a name variable. Do NOT alter it to anything else.

2. Establish associations between phenotypes and genotypes using the novel testing method and return the calculated p-value. Note that genotypes are unphased and allelic status of each locus is encoded by 0, 1, 2 mode, where 0 for wildtype homozygote, 1 for heterozygote and 2 for variant homozygote.

3. At the bottom of the R script, attribute the calculated p-value to a R variable that MUST be named by “pValue” in order for the calculated p-value to be accepted by plot and output functions. E.g.

   pValue <- R_function_to_calculate_p_value(geno, pheno, otherArgs)

3.4.3.4 Output figures and files

Click “Set Path for Output File” to change path, otherwise the current working directory will be used.

In order to draw plots R must be in the system PATH. Once the simulation finishes it outputs two figures to show comparison of selected association methods, where one draws QQ plots for evaluating type I errors and the other shows a bar graph to display power comparison. Results will also be saved into text files.
3.5 Case-control sample rare variant association analysis - an example to use SimRare

This example shows how to use SimRare to evaluate rare variant association methods on simulated case-control samples. It generates the variant data (haplotype) pool using the demographic model based on Kryukov et al. (2009) \cite{72} and phenotype using the case-control prevalence model of a sample (Risch, 1990) \cite{99}.

We generate data under both null ($H_0$) and alternative ($H_1$) hypotheses, respectively, and evaluate association methods for each of them. In the rare variant association tests, $H_0$ assumes that there is no relationship between rare variants and the complex (disease) trait, and $H_1$ vice versa.

3.5.1 Create variant data pool

We first create a variant pool with 50 replicates (50 independent realizations of gene sequence evolution given the same demographic parameters as input) by the following steps.

1. Launch SimRare and activate the module by clicking “Need generate rare variant data pool? –> Click Here”.

2. Select “Use Default Input Parameters?” (default inputs are based on Kryukov et al.’s \cite{72} demographic model).

3. Change “Number of Replicates” to 50.

4. Click “Path” select a desired folder where the output files will be saved, then click “Run”.

88
Refer to section 3.4.1 for the comprehensive meaning of each input parameter shown on the interface.

A snapshot of the interface now should look like Figure 3.4. The outputs files will contain simVarPool.maf, simVarPool.sel and simVarPool.pos, which are the created variant (haplotype) pool for 50 replicates.

3.5.2 Specify disease model parameters under $H_1$

In this example, we focus on the case-control prevalence model of a sample, with the sample size of 2,000 individuals (1,000 cases and 1,000 controls). We specify disease model parameters by the following steps.

1. On SimRare main module choose “Case-control prevalence model of a sample” from “How do you want to establish genotype-phenotype associations?”.

2. Select “Use Default Input Parameters?”.

3. Click “Load .maf file” and select the simVarPool.maf file generated above (SimRare will automatically load .sel and .pos files given that they are saved under the same folder by the same prefix).
Figure 3.5: Specify case-control disease model parameters under the alternative hypothesis ($H_1$)

Refer to section 3.4.2 for the comprehensive list of parameters shown on the interface and their meanings.

4. Change both “# Cases” and “# Controls” to be 1000.

5. Change “Odds Ratio for Detrimental Mutations” to 1.6 and “Odds Ratio for Protective Mutations” to 0.9.

6. Click “Set Path for Output File” and choose a desired folder to save output files (can be the same folder selected in section 3.5.1).

7. Select “Save Input Parameters?”, specify a file name, e.g. *MySimu_H1.par* to save current user inputs, press Enter and then click “Save”. Up to this step a snapshot of the interface should be like that shown in Figure 3.5.
Note. There is no need to run anything with this module since our ultimate goal is to evaluate and compare association methods in section 3.5.4. Thus, it is sufficient here to create a *par* (configuration) file only.

### 3.5.3 Specify disease model parameters under $H_0$

We repeat the steps shown in section 3.5.2 and input 1.0 to both “Odds Ratio for Detrimental Mutations” and “Odds Ratio for Protective Mutations” here, then specify a different configuration file name, *MySimu\_H0.par*, and save the current inputs.

Note. Under the null hypothesis ($H_0$), odds ratio is equal to 1.0 for any type of variants (common, detrimental or protective).

### 3.5.4 Perform association tests

Association tests are performed using parameters settings under $H_0$ and $H_1$, respectively. We may choose to evaluate any one or more test methods. In this example we select CMC-one, WSS-one, KBAC-one and MZ-one and proceed by the following steps.

1. Activate the module of association tests by clicking “Need apply existing/novel association methods? —> Click Here”.

2. Click “Load Configuration file” and select the *MySimu\_H0.par* file.

3. Select existing association methods “CMC-one, MZ-one, KBAC-one and WSS-one”.
4. By default, number of replicates for power calculation is 2000, we change it to 1000 to generate less number of replicates for this example.

5. Click “Set path to output file” to specify a desired folder where output figures will be saved (it can be the same folder selected for sections 3.5.1 and 3.5.2).

6. Specify an output file name (prefix), such as test_H0, then click “Run”. A snapshot of the interface is like Figure 3.6.

**Note.** Running this step requires relatively long time particularly if number of replicates and (or) number of permutations are large. Additionally, it is optional to save data sets of generated genotype/phenotype for all replicates into external files (*_rep1.ped, *_rep2.ped, ..., *_repx.ped) by clicking “Output geno/pheno info for all replicates”. This is not recommended since it can slow down the running speed while creating a large number of files and writing them to the local folder.

7. Repeat steps above but load the other configuration file, MySimu_H1.par.

8. Specify a different file name, test_H1, then click “Run”.

### 3.5.5 Results

SimRare generates graphs of power comparison (Figure 3.7) and QQ plots (e.g. Figure 3.8 for association method “CMC-one”) to evaluate type II and type I errors, respectively.

It is reasonable to obtain that if data is generated under $H_0$, the test power for any association method is around the significance level, which is 0.05 in this exam-
Figure 3.6: Evaluate rare variant association test methods

Figure 3.7: Power comparison of association tests under both $H_0$ and $H_1$

Power comparison of rare variant association tests: CMC-one, MZ-one, KBAC-one and WSS-one, for the case-control prevalence model of a sample. Left - data generated under the null hypothesis ($H_0$); right - data generated under the alternative hypothesis ($H_1$).
Figure 3.8: Quantile-Quantile plots of CMC-one method under both $H_0$ and $H_1$

QQ plots of rare variant association method, CMC-one, with data generated under $H_0$ (left) right, and $H_1$ (right).

Meanwhile, if data is generated under $H_1$ based on the input odds ratios, all methods have approximately equal power to detect associations (about .28 to .30 probability of making type II error), while CMC and WS are slightly better than the others in this particular scenario of study design.

Also, we can view QQ plots for all association methods. As an example Figure 3.8 shows QQ plots for “CMC-one” method. Under $H_0$ the distribution of observed p-values does not inflate compared to the uniform distribution of expected p-values, which demonstrates that “CMC-one” method controls type I error well. In contrast, under $H_1$ the observed significant p-values confirm that the method is capable of detecting associations between genotype and phenotype.

3.6 Power comparison of existing complex trait rare variant association methods

For rare variants power analysis, the power of tests depends not only on sample
size, disease prevalence, minor allele frequency, genetic model and variant effects, but also on the allelic architecture, e.g. the cumulative variant frequency within analyzed region, proportion of causal variants, effects of variants in the same or different directions (protective vs. detrimental).

Using SimRare we generate data under a couple of different disease etiology scenarios using the case-control sample prevalence and population attributable risk models, and evaluate a variety of rare variant association methods for each scenario. With both variant and phenotype data being fairly generated by the unified and realistic framework, we demonstrate that there is not a single most powerful method and most rare variant association methods are not robust.

### 3.6.1 Simulation setting

The variant data are simulated based on the demographic history of European population (Boyko et al., 2008) [15]. The change of effective population size begins with a burn-in stage of 7,947 individuals, followed by a bottleneck stage of 262 individuals for 84 generation, followed by an expansion to 7,019 individuals for 5,217 generations and ends in a second expansion to 52,907 individuals for 576 generation. Other key parameter settings include that mutation rate equals to $1.8 \times 10^{-8}$ per site per generation, gene length is fixed at 1,500 bp (coding region and only variants with frequency of $\leq 1\%$ are analyzed) and nonsynonymous variants within the gene region are analyzed.

The phenotype data (affection status for case-control study) are generated conditional on the multilocus genotype. The parameter setting is determined by the underlying disease etiology. We show power analysis comparison of association
methods below on a number of different scenarios to generate phenotypes.

For the evaluation of power empirically we set significance level (p-value) to be 0.05, number of permutations to be 3,000 and number of replicates to be 2,000. We implement the following test methods to detect associations between genotype and phenotype in the case-control prevalence model of a sample: ANRV/MZ (Morris and Zeggini, 2010) [83], aSum (Han and Pan, 2010) [54], cAlpha (Neale et al., 2011) [85], CMC (Li and Leal, 2008) [76], KBAC (Liu and Leal, 2010) [77], RareCover (Bhatia et al., 2010) [12], RVE (Cohen et al., 2004) [24], SKAT (Wu et al., 2011) [122], TestRare/PRVT (Ionita-Laza et al., 2011) [61], VT (Price et al., 2010) [95] and WSS (Madsen and Browning, 2009) [80].

3.6.2 Case-control prevalence model

For the case-control prevalence model of a sample, Figure 3.9 shows the power comparison with impact of non-causal variants. To generate phenotype data, we assume fixed effect model; odds ratio = 2.0; additive mode of inheritance; disease prevalence = 1%; sample size of 1,000 cases and 1,000 controls with proportion of non-causal variants ranging from 0% to 75%. In another scenario, both detrimental and protective variants are generated within a gene region for a variable effect model, odds ratio ranges from 1.5 to 5 for detrimental variants and from 0.6 to 0.2 for protective variants. Simulation results are shown in Figure 3.10.

3.6.3 Case-control population attributable risk model

Additionally, we perform power comparison by implementing the case-control population attributable risk model of a sample (Madsen and Browning, 2009) [80].
Figure 3.9: Power comparison with impact of non-causal variants

Impact of non-causal variants, fixed effect model OR = 2. Prevalence 1%. N = 1000:1000

Power comparison with proportions of non-causal variants being 0%, 25%, 50% and 75%, respectively. Rare variant association methods are applied on simulated gene sequences of 1,500 bp in the case-control prevalence model of a sample. Variant data are generated by demographic model of the European population (Boyko et al., 2008) [15] and mutation rate = $1.8 \times 10^{-8}$ per site per generation. Sample size = 1,000 cases and 1,000 controls; odds ratio = 2 in the fixed effect model; mode of inheritance is additive; disease prevalence = 1%.
Power comparison with proportions of protective to detrimental variants being 0, 0.25/0.75, 0.5/0.5, 0.75/0.25 and 1.0/0, respectively. Rare variant association methods are applied on simulated gene sequences of 1,500 bp, with both detrimental and protective variants generated in the same gene region for the case-control prevalence model of a sample. Variant data are generated by demographic model of the European population (Boyko et al., 2008) [15] and mutation rate = $1.8 \times 10^{-8}$ per site per generation. Sample size = 1,000 cases and 1,000 controls; odds ratio ranges from 1.5 to 5 for detrimental variants and from 0.6 to 0.2 for protective variants in the variable effect model; mode of inheritance is additive; disease prevalence = 1%.
Power comparison with group population attributable risk being 0.5%, 1%, 1.5%, 2.5% and 5%, respectively. Rare variant association methods are applied on simulated gene sequences of 1,500 bp in the case-control population attributable risk model of a sample. Variant data are generated by demographic model of the European population (Boyko et al., 2008) and mutation rate = $1.8 \times 10^{-8}$ per site per generation. Sample size = 800 cases and 800 controls, mode of inheritance is additive.

Under such scenario, we assume additive mode of inheritance; sample size of 800 cases and 800 controls with group population attributable risk ranging from 0.5% to 5%. Figure 3.11 depicts the power comparison results.

### 3.6.4 Results

From Figures 3.9, 3.10 and 3.11 we observe that various tests have superior power in different circumstances. However, for many tests there is only small differences in their power, e.g. WSS, VT, KBAC and CMC. One-sided tests can increase power
if it is testing for an association with detrimental variants. Most tests that were de-
veloped to detect associations with variants within a region which are bidirectional
are not very powerful when the variant effects are unidirectional.

Besides, confounding can also lead to spurious associations, e.g. population sub-
structure. Therefore, in addition to using power as a criterion to choose a test,
the ability of a test to control for potential confounders should be an additional
consideration.
Chapter 4

Modeling of adult hippocampal neurogenesis in its early stages

4.1 Motivation and overview

Adult hippocampal neurogenesis, a process of formation of new neurons, occurs throughout life in the hippocampus, an area responsible for learning and memory. The majority of the hippocampal neurogenic studies have predominantly focused on the late stages, particularly on the role of newborn neurons but little is known about its early stages that regulate the proliferation and differentiation of neural stem cells and progenitor cells to immature neurons. In this chapter, based on the branching process theory and biological evidence, we develop a computational model that represents the early-stage hippocampal neurogenic cascade and allows us to predict the overall efficiency of hippocampal neurogenesis in both normal and diseased conditions. Using the model we derive the equilibrium distribution of cell population and simulate the progression of cell labeling intensity for a BrdU pulse-and-chase labeling scheme to fit the model and simulation results in the experimental data. Genetic algorithm is adapted in this study as the parameter searching heuristic. Our simulation results reveal unknown but meaningful biological param-
eters, among which the most crucial ones are apoptotic rates at different cell stages. Our study concludes that apoptotic rates reach maximum for neuroblasts. Renewal probability of neural progenitor cells is low; variance of cell durations of neuroblasts is high and expected duration for apoptotic cells is short.

4.2 Introduction

Adult neurogenesis, the formation of new neurons throughout life, is a phenomenon very limited in mammals and apparently reduced to two distinct regions of the brain: the subventricular zone (SVZ), involved in olfactory processes, and the sub-granular zone (SGZ) of the dentate gyrus (DG), where the new neurons seems to be relevant not only for learning and memory formation but for mood disorders as well (for review see Ref Deng et al., 2010) [32]. In the SGZ, a population of radial glia-like cells, described also as type-1 cells or quiescent neural progenitors (QNP) (Kempermann et al., 2004) [64], gives rise to neuroprogenitors cells (NPCs) which eventually mature and differentiate into functional granular cells (GC) (Zhao et al., 2006) [126]. However, the majority of NPCs undergo death by apoptosis in the first 1 to 4 days (first critical period of survival), during the transition from amplifying neuroprogenitors (ANP) to neuroblasts (NB) and only a small subset integrates into the hippocampal circuitry as mature neurons at the end of a 4-5 week period (Sierra et al., 2010) [109].

Hippocampal neurogenesis can be regulated both positively and negatively by external stimuli, such as learning (Gould et al., 1999) [52], exercise (van Praag et al., 1999) [113], environment (Kempermann et al., 1997) [63] and stress (Revest et al.,
The functionality of the new GCs formed has been studied widely. During the last years, behavioral experiments involving new transgenic models have shown the important role of hippocampal neurogenesis in spatial relational memory formation (Dupret et al., 2008; Zhang et al., 2008) Many issues regarding the neurogenic cascade are not known because they are experimentally untestable at this time. For example, currently there is no method that would allow us to distinguish each stage of the neural progenitor cell (NPC) at a given moment of neurogenesis, nor to predict how long it will be in that stage; also there is lacking methods to observe cell death rate at each stage. Such information is very important, because specific, targeted modifications of neurogenesis may be beneficial for many disorders such as dementia and depression.

Additionally to in vivo experiments, several computational models of hippocampal neurogenesis have been developed trying to understand the effect of new granular cells incorporation in the DG. These models have shown that new GC participate on pattern separation, avoiding interference between memories while older ones are not greatly disturbed (Aimone et al., 2009; Weisz et al., 2009; Aimone et al., 2010).

However, computational models of the hippocampal adult neurogenesis have predominantly focused on proliferation of stem cells and on late stages of neurogenesis, when immature neurons develop into mature neurons and are incorporated into the local circuitry. Thus, these models have not addressed all the processes that occur throughout the neurogenic cascade, specifically all three possible fates of cells that are part of the neurogenic cascade - proliferation, differentiation and cell death. Our
QNPs (quiescent neural progenitors) provide a basal level influx of new ANPs (amplifying neural progenitors) through asymmetric divisions. Newborn ANP proliferate several times but only some of them can survive to differentiate into the EN (early neuroblast). As EN keep differentiating, their numbers will be reduced further. In the end, only about few granule neuron cells will be produced. The apoptotic cells have short lifetimes. After undergoing apoptosis, they are rapidly phagocytosed and degraded by resident microglia. See details in Sierra et al., (2010) [109].

Experimental results suggest that the number of newborn neurons which eventually reach maturity largely depends on the early stages of neurogenesis, when most of cells die via apoptosis. Figure 4.1 illustrates the schematic way of presenting the neurogenic cascade in the adult hippocampus with apoptosis. Hence, we design a comprehensive computational model of all early stages of neurogenic cascade, including transition, proliferation, differentiation and survival of newborn cells from the stage of a neural progenitor to the stage of a mature neuron.

To determine the number of proliferating cells in the neurogenic system, we use BrdU (Bromodeoxyuridine), which labels cells in S-phase of the cell cycle. BrdU labeling can be performed as a single or cumulative injection. In Figure 4.2 we
Paradigm of single labeling and continuous labeling experiments. Arrows in different paradigms pointing towards green areas indicate time points of BrdU injection for two experimental designs respectively, while those pointing to numbers stand for time points when animals are sacrificed. For the purpose of modeling and computation of this study we focus on the single BrdU labeling paradigm only.

show the two labeling experiments. In single labeling experiments, animals are injected with BrdU at $t = 0$ whereas in continuous labeling experiments animals are treated with one injection every 3 hours in the first 24 hrs, total of 9 injections including the one at time 0. Then animals are sacrificed at different time points, when we analyze the total number of BrdU cells as well as the percentage of cells in each stage of the cascade. The computational part of this study focuses on the single BrdU labeling experiments only.

4.3 Materials and methods

4.3.1 Animals

Wild-type (C57BL/6) or transgenic nestin-CFPnuc (Encinas et al., 2006) [41] mice were used. All mouse studies were approved by the Baylor College of Medicine Institutional Animal Care and Use Committee and performed in accordance with institutional and federal guidelines. Unless otherwise stated, animals were 1 month
old. In Bromodeoxyuridine (BrdU) pulse-and-chase experiments, mice were injected with BrdU (250 mg/kg, i.p.) once (single BrdU paradigm) or every three hours for 24h (cumulative BrdU paradigm), and sacrificed at different time points.

4.3.2 Histology

Mice were transcardially perfused with phosphate buffer saline (PBS) followed by 4% paraformaldehyde (PFA). The brains were dissected out submerged into 4% PFA for 4hr at room temperature (RT) and sectioned using a vibratome. For immunofluorescence, free-floating sections were immunolabeled according to conventional procedures. The brains were dissected out, and then transferred to a cryoprotectant solution (30% sucrose, 30% ethylene glycol in PBS) and stored at -20°C. Once brains were well cryoprotected, six series of 50 µm lateral sections were collected using a vibratome. A full series of free-floating sections were immunostained following conventional procedures (Encinas and Enikolopov, 2008). Briefly, all washes and incubations were done in PBS containing 3% bovine serum albumin (BSA; Sigma-Aldrich) and 0.3% Triton X-100 (Sigma-Aldrich). An antigen retrieval step (2N HCl, 15min, 37°C) for BrdU detection was performed, followed by extensive washes with borate buffer (0.1M). Sections were pre-incubated in PBS containing 3% BSA, 5% normal goat serum (Vector Labs) and 0.3% Triton X-100 for 2h at RT, followed by overnight incubation with primary antibodies (see below) at 4°C. After extensive washing, sections were incubated with the appropriate secondary antibody conjugated with Alexa 488 (Molecular Probes), Rhodamine Red-X and Cyanine 5 (Jackson Immunoresearch) together with DAPI (5µg/mL, Sigma-Aldrich) for 2h at RT. They were then washed, and mounted on
slides with Fluorescent Mounting Medium (Dako). The following primary antibodies were used: BrdU (1:400, Accurate); DCX (Cell Signaling, 1:200); GFAP (1:1,000, Sigma-Aldrich); NeuN (1:1,000, Chemicon).

4.3.3 Confocal Microscopy

Sections were imaged with a Zeiss LSM or a Leica SP5 confocal microscope. The number of apoptotic cells and (or) BrdU+ cells per z-stack was estimated via the optical dissector method (Encinas and Enikolopov, 2008) [12]. Blind analysis was performed with AxioVision 4.5 (Zeiss) or LAS AF Lite (Leica). Two-three 20µm z-stacks (consisting of 30 optical slices of 0.8µm thickness) were obtained from every section. The number of cells was evaluated as a function of the volume of the SGZ, defined as the SGZ length in the image multiplied by an optical thickness of 20µm and a height of 20µm (which we defined in these experiments as a layer of cells expanding 5µm into the hilus and 15µm into the granular layer), then extrapolated to the volume spanned by the SGZ in the hippocampus.

4.3.4 Cell transit times

Cell transit time is defined as the duration time of cell spent in the phase or stage before it transits into the next phase or stage. Same as described in section 5.3.2.3, we use shifted gamma distributions to model the lifetime of cells. Such distribution has three parameters, \((k, s, v)\), with its probability density function given as

\[
 f(x|k, s, v) = (x - v)^{k-1} \frac{e^{-\frac{(x-v)}{s}}}{s I(k)} ,
\]

where \(k\) is the shape parameter, \(s\) is the scale parameter and \(v\) is the shift value (minimum duration). See section 5.3.2.3 for details.
4.3.5 Classification of cell stages

The neurogenic cascade consists of multiple cell types (different states) and probabilities of cells progressing from one type to another (transition rates) between states. Our model includes the following cell types: quiescent QNPs, ANPs, Neuroblasts (including immature neurons), granule neurons and apoptotic cells.

- **QNP (quiescent neural progenitor)**

  QNP cells provide the ultimate influx of new born ANPs which massively proliferate to drive the entire cell population evolve to produce matured granule neuron cells. The majority of QNPs are kept in quiescence while activated ones divide asymmetrically to enrich the pool of new born ANPs. The basal line QNP-ANP influx can be modeled as a homogeneous Poisson Process. Even though the influx rate changes as the animal is aging, we assume it to be fixed as if we froze and took a snapshot of each 1-month-old animal brain while we counted number of BrdU labeled cells in the pulse-and-chase labeling experiments.

  At the beginning of BrdU injection, all activated QNPs that are in S-phase will be labeled. Encinas et al., (2011) [40] indicate that each newly activated QNP proliferates three times to produce three ANP progeny through asymmetric divisions and eventually becomes an astrocyte. Experimental study by Sierra et al., 2010 [109] shows that QNP do not undergo apoptosis.

- **ANP (amplifying neural progenitor)**

  Each new born ANP proliferates several times. After the cell divides for a min-
imum number of times ($\geq 1$) it can either continue proliferating, differentiate or die. We model ANP progression by specifying minimum and maximum numbers of divisions and renewal probability, which is the probability of an ANP continuing to proliferate after finishing its minimum number of divisions. If an ANP chooses to proliferate it enters cell cycle phases of $G_1$, $S$ and $G_2M$, otherwise it becomes a non-proliferating ANP which may choose to differentiate to NB by entering ANP-NB transition state or to undergo programmed cell death by entering ANP-Apop transition state.

**Note.** We assume that cells which are in either ANP-NB or ANP-Apop stage are non-proliferating ANP cells.

- **NB (Neuroblast) and IN (immature neuron)**

Neuroblasts are non-proliferating cells that are differentiated from ANPs. Cells that are in ANP-NB state are destined to become NBs. The cell duration in NB stage is relatively long and of large variation (1d-12d), and at the end of each NB it may choose to become an IN (immature neuron) or enter apoptosis. Like NB, any immature neuron takes a period of duration time and in the end differentiates into a mature neuron (granule cell) or enter programmed cell death (apoptosis).

**Remark.** IN and NB cells are labeled by the same markers (DCX+ and NeuN+) while INs are distinguished from NBs by morphology. Cells during the transit state normally do not display distinctive morphology. To determine whether a cell is a NB or an IN is often subjective and biased.
To avoid such bias, NB and IN stages can be merged as “NB” stage in the broad view for cells that are labeled by both DCX and NeuN.

- **GC (granule neuron cell)**

  GCs are differentiated cells that will stay in dentate gyrus to form neural connection with existing neurons. Once a GC is formed it cannot die or differentiate anymore.

- **ANP-NB**

  Intermediate state of transition time from a non-proliferating ANP to become a NB, where the duration is modeled in the same way as duration of any other cell type is as a shifted-gamma distribution.

- **ANP-Apop**

  Intermediate state of transition time for a non-proliferating ANP to become an apoptotic cell.

- **Apop (apoptotic cell)**

  We assume that apoptosis may occur at the end of any of the following cell stages, ANP-G₁, ANP-S, ANP-G₂M, non-proliferating ANP and NB. At the moment of BrdU injection all cells that are dividing and in S-phase will be labeled. From the observation of apoptotic-BrdU+ cell labeling curve there are no labeled apoptotic cells observed at either 2 hrs or 12 hrs (Sierra et al., 2010) [109] (see Table 4.2). The estimated duration of ANP-G₂M-phase is about 2hrs (Encinas et al., 2011) [10]. A proportion of newly labeled cells that are at their final allowed division and also transiting from S to G₂M phase
can be captured by BrdU. These observations indicate the existence of ANP-
Apop stage, otherwise the apoptotic-BrdU cells should be seen at 2 hrs or 12
hrs after BrdU injection. Also, they imply that the apoptotic rate of cells in
either S or G2M-phase should be small, otherwise, cells that are labeled in late
S-phases can enter cell death immediately after BrdU injection.

4.3.6 Expected number of ANP divisions

We denote $a$ and $b$ as the minimum and maximum number of divisions of each
newborn ANP, where $a$ is the required minimum number of divisions and $b$ is
the maximum allowed number of divisions. We further denote $p$ as the renewal
probability of each ANP (probability of proliferating after dividing $a$ times) and
denote $X$ as the random variable of number of progeny produced by each new born
ANP. Therefore, we obtain

\[ P(X = 2^a) = 1 - p, \quad P(X = 2^{a+i}) = p^i(1 - p), \quad \text{for } 1 \leq i \leq b - a - 1 \text{ and} \]
\[ P(X = 2^b) = p^{b-a}. \]
\[ E(X) = 2^a(1 - p) + \sum_{i=1}^{b-a-1} p^i(1 - p)2^{a+i} + 2^b p^{b-a} \]

For $a < b$, the expected number of ANP divisions can be derived as $log_2E(X)$.
If $a = b$, the expected number of division is $a$.

4.3.7 Modeling of the neurogenic cascade using branching process

Overall, we expect to build a model for neurogenesis in early stages from QNP to
newborn neurons with consideration of factors focused on the ANP fate selection
and cell death rates. We use the branching processes, which consider behavior
of abstract particles, subject to certain independence assumptions. Particularly,
Figure 4.3: Branching process modeling of the neurogenic cascade

Hierarchical structure of the neurogenic cascade modeled by Branching process with different cell types as different compartments. $\lambda =$ intensity of influx of new ANPs from QNPs, $\times 2 =$ cell doubles, $d_i =$ cell death rate at the end of stage $i$, $p =$ renewal probability of ANPs, $T_i =$ duration time of cell in stage $i$.

the proliferation of activated neurogenic progenitors is a process, which requires description in terms of multitype branching process. Thus, we choose Multitype Bellman-Harris branching process to model the neurogenic cell population and resulting labeling curves. Besides the fact that Bellman-Harris process is frequently employed to model proliferation of systems of biological cells (Kimmel and Axelrod, 2002) [65], it is necessary to distinguish between cells in different cell cycle phases other than only the hierarchical nature of the process (Yakovlev and Yanev, 2006) [123].

The structure of the model is presented in Figure 4.3, which is constructed based on our experimental observations (Sierra et al., 2010) [109].

We model the hierarchical structure with transit probabilities of cells from any stage to the next possible one. Thus, cell death rates for different cell phases/stages are modeled by corresponded transit probabilities into apoptosis, where symbols $d_i$ denotes the cell death rate of cell type $i$. Furthermore, if we let $p$ be the renewal probability of ANP, by the law of total probability, all the other transit rates through different compartments of the model may be specified with certain probabilities.
Consider a collection of proliferating particles of \( I \) types, which proliferate according to the following rules:

- At time \( t = 0 \), an ancestor particle of type \( i \) is born, which lives for a random time \( \tau \) with cumulative distribution function (cdf) \( T_i \) and upon death, produces a random number of progeny of all types, described by a vector \((X_1, \ldots, X_I)\) with multivariate probability generating function \( h_i(s_1, \ldots, s_I) \).

- At time \( t = \tau \), each first-generation progeny particle of type \( j \) lives for a random time with cumulative distribution function (cdf) \( T_j \) and upon death, produces a random number of progeny of all types, described by vector of multivariate pgf \( h_j(s_1, \ldots, s_I) \), independently of other progeny particles.

- The cycle of life, death and progeny production is repeated indefinitely by each generation of particles.

If we denote the multivariate pgf of number of particles of all types present in the process initiated by an ancestor of type \( i \) with \( F_i(s, t) \), we obtain the Bellman-Harris integral equation for this scenario as

\[
F_i(s, t) = \int_0^t h_i[F(s, t - \tau)]d_\tau T_i(\tau) + s_i[1 - T_i(t)]
\]

Differentiating both sides of the equation above with respect to \( s_j \) and setting \( s_1 = s_2 = \ldots = s_I = 1 \), we obtain the following equation for the matrix \( M(t) = [M_{ij}(t)] \), where \( M_{ij}(t) \) is the expected number of particles of type \( j \) at time \( t \), in the process.
initiated by the ancestor particle of type $i$ at time 0

$$M_{ij}(t) = \int_0^t \sum_{k=1}^I m_{ik}M_{kj}(t-\tau)d\tau + \delta_{ij}[1 - T_i(t)]$$

where $m_{ij}(t)$ is the expected number of progeny of type $j$ of a particle of type $i$, $\delta_{ij} = 1$, if $i = j$, otherwise, $\delta_{ij} = 0$. With $*$ representing the convolution notation, the above equation can be expressed as

$$M_{ij}(t) = \sum_{k=1}^I m_{ik}M_{kj}(t) * T_i(t) + \delta_{ij}[1 - T_i(t)]$$

In matrix notation, we obtain

$$M(t) = T(t) * [mM(t)] + [I - T(t)]$$

where $I$ is the identity matrix and $G = diag(G_1, ..., G_I)$. This is an equation of the renewal process, which has a unique solution of locally bounded variation if $G(0) = 0$, expressed by the infinite series

$$M = \sum_{k=0}^\infty (Tm)^*k * (I - T) \tag{4.1}$$

which yields the fundamental solution of the mathematical modeling of the neurogenic cascade as a matrix function of time for the number of cells of each type, where $M_{ij}(t)$ is the expected number of cells of type $j$ at time $t$, in the process initiated by an ancestor cell of type $i$ at time 0, $T = diag(T_1, ..., T_I)$ denotes the diagonal matrix with the distribution function of lifetime (or duration) $T_i$ of each cell $i$, $m$ is the transition matrix and $m_{ij}(t)$ is the expected number of progeny of
cell of type $j$ produced by a cell of type $i$, obtained from the multivariate pgf of numbers of progeny produced by the type $i$ cell, and $I$ is the identity matrix.

In addition, symbol $\ast$ represents the convolution notation, which is defined by the following Lebesgue-Stieltjes integral if $A(t)$ and $B(t)$ are two right-continuous functions with locally bounded variation on $[0, \infty)$

$$A(t) \ast B(t) = \int_0^t A(t - \tau)d\tau B(\tau)$$

Based on the model structure, we have the transition matrix $m$ as (e.g. if minimum and maximum numbers of ANP divisions are 1 and 3)

$$
\begin{array}{cccccccccccc}
G_1 & S & G_2M & G_1 & S & G_2M & G_1 & S & G_2M & A - N & NB & A - A & Apop \\
\hline
G_1(1) & 0 & \overline{G}_{G_1} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & d_{G_1} \\
S(1) & 0 & 0 & \overline{G}_{S} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & d_{S} \\
G_2M(1) & 0 & 0 & 0 & 2\overline{G}_{G_2M} & 0 & 0 & 0 & 0 & 0 & x^* & 0 & y^{**} & d_{G_2M} \\
G_1(2) & 0 & 0 & 0 & 0 & \overline{G}_{G_1} & 0 & 0 & 0 & 0 & 0 & 0 & d_{G_1} \\
S(2) & 0 & 0 & 0 & 0 & \overline{G}_{S} & 0 & 0 & 0 & 0 & 0 & 0 & d_{S} \\
G_2M(2) & 0 & 0 & 0 & 0 & 0 & 2\overline{G}_{G_2M} & 0 & 0 & x^* & 0 & y^{**} & d_{G_2M} \\
G_1(3) & 0 & 0 & 0 & 0 & 0 & \overline{G}_{G_1} & 0 & 0 & 0 & 0 & d_{G_1} \\
S(3) & 0 & 0 & 0 & 0 & 0 & \overline{G}_{S} & 0 & 0 & 0 & 0 & d_{S} \\
G_2M(3) & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & z^# & 0 & w^{##} & d_{G_2M} \\
A - N & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\
NB & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & d_{NB} \\
A - A & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\
Apop & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
\end{array}
$$

where A-N represents ANP-NB stage, A-A is for ANP-Apop stage, $\overline{G}_i = 1 - d_i$,

$$x^* = 2\overline{d}_{G_2M}(1 - p)\overline{d}_{ANP}, \quad y^{**} = 2\overline{d}_{G_2M}(1 - p)d_{ANP}, \quad z^# = 2\overline{d}_{G_2M}\overline{d}_{ANP} \quad \text{and} \quad w^{##} = 2\overline{d}_{G_2M}d_{ANP} \quad (d_{ANP} \text{ is the cell death rate of non-proliferating ANPs}).$$

Furthermore, to model the QNP to ANP influx we assume that any ‘arrival’ of a
new ANP is independent of all previous ‘arrivals’ and number of new ANPs arrived during a period of time is depends only on the length of that period times the intensity of the influx, $\lambda$. Thus, such process is a Poisson process with intensity parameter $\lambda$, and the probability that the number of new ANPs arrived during a time unit ($u$) being equal to $n$ can be derived as

$$P[N(t + u) - N(t) = n] = \frac{e^{-\lambda u} (\lambda u)^n}{n!}, n = 0, 1, ...$$

where $N(t)$ is the number of new ANPs arrived until time $t$ and $N(t + u)$ is the number of new ANPs arrived until time $t + u$.

### 4.3.8 Modeling and simulation of cell labeling curves

Our data is acquired by time-course labeling experiments of the single BrdU injection with curves of total and partial counts at different times of measurements. Given the estimated number of cells during the S-phase in each stage at the beginning of BrdU injection, we may calculate the number of labeled cells of each type at any moment by equation (4.1). However, solving it in analytical forms is so complicated (see last paragraph of section 4.4.3 for detail) that simulation, a more convenient computational tool, need to be applied.

An alternative approach to computationally produce labeling curves is the event-based simulation method. Assuming that we have computed the numbers of cells in different stages at the moment of BrdU injection ($t = 0$), we can trace the fate of labeled cells at unit time points by recording their behaviors. Briefly, a series of random numbers are generated for the probability of time for which a labeled
cell will stay in the current stage and the probability of that cell to transit to the next stage, until the cell enters apoptosis or becomes a matured neuron. The probabilistic path is repeatedly generated for a large number of times until each labeled cell in the population has been simulated. Beginning to trace the entire process from $t = 0$, we can rebuild the labeling curves in silico by accumulating the fates of all labeled cells up to particular moments (e.g. times of measurements in experiments).

A simulation program carrying out tasks outlined above, has been written in Python programming language. This program can both compute distribution of initial cell population and generate BrdU labeling curves. Figure 4.4 illustrates the event-based simulation scheme of the dynamic of a neural progenitor cell.

4.3.9 Parameter search and goodness-of-fit

We search input parameter space to discover parameter combinations that can best fit the experimental data. We adapt genetic algorithm as the searching heuristic since the parameter space is too complex to be applicably represented by enumeration of all possible combinations. Refer to section 1.5.3 for detail about the
Table 4.1: Model parameters that can be estimated experimentally or not

<table>
<thead>
<tr>
<th>Experimentally estimable</th>
<th>Difficult to determine by experiments</th>
</tr>
</thead>
<tbody>
<tr>
<td>average duration times</td>
<td>intensity of QNP→ANP influx</td>
</tr>
<tr>
<td>cell population size</td>
<td>apoptotic rate at each stage</td>
</tr>
<tr>
<td></td>
<td>ANP renewal probability</td>
</tr>
<tr>
<td></td>
<td>possible number of ANP divisions</td>
</tr>
<tr>
<td></td>
<td>shapes of distributions of durations</td>
</tr>
<tr>
<td></td>
<td>minimum durations</td>
</tr>
</tbody>
</table>

algorithm.

During the search any parameter set is evaluated by computing variance weighted least square error $\sum \frac{(E-S)^2}{\sigma^2}$ (since variances of measurements on different time points are non-constant) to test how well the simulated results fit to data of labeling curves. $E$ and $\sigma^2$ are mean and variance of experimental data at a given time point, whereas $S$ is the corresponded simulation result. $\sum$ sums over all available time points of measurements.

A brief list of model parameters is shown in Table 4.1, some of them can be experimentally estimable, whereas most of them can not be determined experimentally but represent valuable biological meanings.

4.4 Results and discussion

4.4.1 Data

Our data were obtained by \textit{in vivo} BrdU pulse-and-chase labeling experiments over a period of 32 days from 1 month old mice brains.

In experiment 1 (Sierra \textit{et al.}, 2010) \cite{109}, number of total BrdU positive cells and number of BrdU+ apoptotic cells were counted at 12 different time points of measurements ($t = 2hr, 12hr, 1d, 2d, 3d, 4d, 8d, 11d, 15d, 18d, 22d, 32d$, assuming
Table 4.2: Data of total BrdU+ cell count and BrdU+ apoptotic cell count

<table>
<thead>
<tr>
<th>Time (days)</th>
<th>Experiment 1</th>
<th>n</th>
<th>Total BrdU+ cells</th>
<th>BrdU+ apoptotic cells</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.08 (2hr)</td>
<td>3</td>
<td>2690 (320)</td>
<td>0 (0)</td>
<td></td>
</tr>
<tr>
<td>0.5 (12h)</td>
<td>2</td>
<td>4157 (784)</td>
<td>0 (0)</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>4</td>
<td>5392 (557)</td>
<td>40 (18)</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>5</td>
<td>5803 (138)</td>
<td>121 (33)</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>3</td>
<td>4781 (344)</td>
<td>48 (25)</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>5</td>
<td>4186 (201)</td>
<td>23 (14)</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>6</td>
<td>3518 (307)</td>
<td>10 (11)</td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>3</td>
<td>2427 (202)</td>
<td>0 (0)</td>
<td></td>
</tr>
<tr>
<td>15</td>
<td>4</td>
<td>1342 (185)</td>
<td>33 (13)</td>
<td></td>
</tr>
<tr>
<td>18</td>
<td>3</td>
<td>1233 (302)</td>
<td>0 (0)</td>
<td></td>
</tr>
<tr>
<td>22</td>
<td>4</td>
<td>752 (53)</td>
<td>0 (0)</td>
<td></td>
</tr>
<tr>
<td>32</td>
<td>3</td>
<td>950 (234)</td>
<td>13 (16)</td>
<td></td>
</tr>
</tbody>
</table>

n - sample size, number of animals sacrificed. Values shown are the mean and standard error of the mean (sem) in number of cells.

that BrdU injection occurred at \( t = 0 \). For each measurement a sample of 2-5 mice were sacrificed. See Table 4.2 for data of counts of total BrdU+ and BrdU+ apoptotic cells over time.

Additionally, two independent experiments were carried out to estimate the proportion of each type of BrdU+ cells (QNP, ANP, NB and GC) at several time points from randomly selected areas of animal brains. Experiment 2 used QNP and ANP specific markers (GFP and GFAP) with BrdU to count BrdU+ QNP and BrdU+ ANP cells at \( t = 2\, hr, 1\, d, 2\, d, 4\, d \) and \( 8\, d \). Experiment 3 employed NB and GC specific markers (DCX and NeuN) with BrdU to calculate BrdU+ NB and BrdU+ GC cells at \( t = 1\, d, 2\, d, 4\, d, 8\, d, 15\, d \) and \( 32\, d \). Specifically, QNPs are labeled by GFP and GFAP, ANPs are labeled by GFP but not GFAP, NBs are labeled by both NeuN and DCX and GC are labeled by NeuN but not DCX. Table 4.3 summarizes results of experiments 2 and 3.
Table 4.3: Data of estimated proportion of BrdU+ cell of each type

<table>
<thead>
<tr>
<th>Time (days)</th>
<th>Experiment 2</th>
<th>Experiment 3</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>n</td>
<td>ANP</td>
</tr>
<tr>
<td>0.08 (2hr)</td>
<td>4</td>
<td>11.16 (2.14)</td>
</tr>
<tr>
<td>0.5 (12h)</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>1</td>
<td>4</td>
<td>5.68 (0.57)</td>
</tr>
<tr>
<td>2</td>
<td>5</td>
<td>3.29 (0.79)</td>
</tr>
<tr>
<td>3</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>4</td>
<td>5</td>
<td>2.53 (0.69)</td>
</tr>
<tr>
<td>8</td>
<td>5</td>
<td>0 (0)</td>
</tr>
<tr>
<td>11</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>15</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>18</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>22</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>32</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

*n* is the sample size (number of animals sacrificed), "-" means no available data. Two groups of animals (all 1 month old) were used for experiments 2 and 3. Values shown are the mean and standard error of the mean (sem) in proportion (×100) of cells of each type.

### 4.4.2 Data cleaning and transformation

#### 4.4.2.1 Adjust cell proportions

At time points \( t = 1\text{d}, 2\text{d}, 4\text{d} \) and \( 8\text{d} \) the sum of observed proportions of all cell types (QNP, ANP, NB, GC and Apoptotic cells, see Tables 4.3 and 4.2) is consistently greater than 1. This discrepancy is due to the nonspecific labeling or identification during transitions from ANPs to NBs. Cells that are in the intermediate stage can be labeled by both ANP and NB markers.

We assume that the estimated proportions on QNP and GC cells are realistic while those on ANP and NB cells at middle time points are inflated due to nonspecific labeling so that they need to be adjusted. At early (2hr) or late time points (15d and 32d) we keep original data since except QNP and GC the labeled cells are either ANP or NB exclusively.

For any time point \( t \) \((t = 1\text{d}, 2\text{d}, 4\text{d} \text{ or } 8\text{d})\), we adjust the sum of proportions of
all types of BrdU+ cells (QNP, ANP, NB, GC and Apoptotic cells) to be equal to 1 (impact of BrdU+ astrocytes is little and can be neglected since the proportion of BrdU+ astrocytes is very small when \( t \leq 8 \) days (Encinas et al., 2011)) \[40\].

The proportion of excessive amount of cells (denoted by \( d_t \)) between ANP and NB is equal to the sum of observed proportions minus 1. We denote \( \alpha_t \) as the ratio that \( \alpha_t d_t \) is the proportion of double labeled cells that belong to ANPs, whereas \((1 - \alpha_t)d_t\) to NBs.

Let \( X = (X_1, X_2)^T \) and \( X' = (X'_1, X'_2)^T \) as two random vectors to represent the proportions of ANP and NB cells before and after transformation, respectively. Therefore, we obtain

\[
X'_1 = X_1 - (1 - \alpha_t)(X_1 + X_2 - d_t) = \alpha_t X_1 + (\alpha_t - 1)X_2 + (1 - \alpha_t)d_t
\]

\[
X'_2 = X_2 - \alpha_t(X_1 + X_2 - d_t) = -\alpha_t X_1 + (1 - \alpha_t)X_2 + \alpha_t d_t
\]

and

\[
X' = A \ast X + B
\]

where

\[
A = \begin{pmatrix}
\alpha_t & \alpha_t - 1 \\
-\alpha_t & 1 - \alpha_t
\end{pmatrix},
B = \begin{pmatrix}
(1 - \alpha_t)d_t \\
\alpha_t
\end{pmatrix}.
\]

Thus, we can compute the mean and variance of \( X' \) as

\[
E[X'] = (\alpha E[X_1] + (\alpha_t - 1)E[X_2] + (1 - \alpha_t)d_t, -\alpha_t E[X_1] + (1 - \alpha_t)E[X_2] + \alpha_t d_t)
\]

\[
\Sigma[X'] = \begin{pmatrix}
V[X'_1] & COV[X'_1, X'_2] \\
COV[X'_1, X'_2] & V[X'_2]
\end{pmatrix} = A^T \ast \Sigma[X] \ast A
\]
Table 4.4: Cleaned data of estimated proportion of BrdU+ cell of each type

| Time (days) | Experiment 2 | | Experiment 3 | |
|-------------|-------------|------------------------------|------------------------------|
| n           | QNP         | ANP                          | n                           | NB | GC |
| 0.08 (2hr)  | 4           | 11.16 (2.14)                 | -                           | -  | -  |
| 0.5 (12h)   | -           | -                            | -                           | -  | -  |
| 1           | 4           | 5.68 (0.57)                  | 3                           | 42.59 (3.95) | 0.2 (0.24) |
| 2           | 5           | 3.29 (0.79)                  | 2                           | 64.41 (5.08) | 0.32 (0.46) |
| 3           | -           | -                            | -                           | -  | -  |
| 4           | 5           | 2.53 (0.69)                  | 3                           | 85.05 (0.82) | 1.52 (0.51) |
| 8           | 5           | 0 (0)                        | 3                           | 93.11 (1.05) | 2.48 (0.31) |
| 11          | -           | -                            | -                           | -  | -  |
| 15          | -           | -                            | 2                           | 86.61 (1.26) | 4.72 (0.05) |
| 18          | -           | -                            | -                           | -  | -  |
| 22          | -           | -                            | -                           | -  | -  |
| 32          | -           | -                            | 3                           | 14.86 (3.62) | 77.34 (6.81) |

n is the sample size (number of animals sacrificed), “-” means no available data. Numbers in bold are adjusted values of proportions (×100). Values shown are the mean and standard error of the mean (sem) of the proportion (×100) of each type of cells.

where $\Sigma[X']$ and $\Sigma[X]$ are covariance matrices of $X'$ and $X$ with $\Sigma[X] = \begin{pmatrix} V[X_1] & 0 \\ 0 & V[X_2] \end{pmatrix}$

by assuming that $COV[X_1, X_2] = 0$.

Using equations above we calculate the adjusted estimates of means and sems of ANP and NB cell proportions, see Table 4.4

**Note.** For any interval time point $t$ ($t = 1d, 2d, 4d, 8d$), $\alpha_t$ is assumed to be $1/2$ since there is no prior knowledge about it.

4.4.2.2 **Transform proportions to cell numbers**

For searching parameters that can best fit the data, total number of BrdU+ cells and estimated number of BrdU+ cells in each specific type are required to evaluate the goodness-of-fit. Non-zero data points expressed as proportions (random variables) from experiments 2 and 3 (Table 4.3) are transformed back to cell counts with re-calculated means and variances.
For any time point \( t \), we assume that the count of total number of BrdU+ cells in an animal is a normally distributed random variable \( Y \sim N(\mu_Y, \sigma_Y^2) \). Thus, for a size of \( n_Y \) samples, we obtain that \( \hat{\mu}_Y = \bar{Y} \) and \( \hat{\sigma}_Y^2 = S_Y^2 \) (data in Table 4.2 are used to estimate \( \hat{\mu}_Y \) and \( \hat{\sigma}_Y^2 \)).

For any specific cell type \( i \) (e.g. ANP), if we denote \( X \) as the count of number of BrdU+ type \( i \) cells at time \( t \), we have

\[
X|Y, P \sim binomial(Y, P)
\]

where we assume that the observed proportion of type \( i \) cell, \( P \), is Gaussian distributed, s.t. \( P \sim N(\mu_P, \sigma_P^2) \). For a size of \( n_P \) samples, \( \hat{\mu}_P = \bar{P} \) and \( \hat{\sigma}_P^2 = S_P^2 \) (\( \hat{\mu}_P \) and \( \hat{\sigma}_P^2 \) are estimated from data shown in Table 4.4).

Assuming that \( Y \) and \( P \) are independent random variables, we can derive that

\[
\]

and

\[
V[X] = V[E[X|Y, P]] + E[V[X|Y, P]]
\]

\[
= V[YP] + E[YP(1 - P)]
\]

\[
\]

\[
\]

Since

\[
\]
Table 4.5: Summary of estimated cell counts in each type

<table>
<thead>
<tr>
<th>Time (days)</th>
<th>total</th>
<th>Apop</th>
<th>QNP</th>
<th>ANP</th>
<th>NB</th>
<th>GC</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.08 (2hr)</td>
<td>2690 (320)</td>
<td>0 (0)</td>
<td>300 (81)</td>
<td>2288 (298)</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>0.5 (12h)</td>
<td>4157 (784)</td>
<td>0 (0)</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>1</td>
<td>5392 (557)</td>
<td>40 (18)</td>
<td>306 (46)</td>
<td>2738 (334)</td>
<td>2296 (297)</td>
<td>11 (11)</td>
</tr>
<tr>
<td>2</td>
<td>5803 (138)</td>
<td>121 (33)</td>
<td>191 (47)</td>
<td>1758 (154)</td>
<td>3738 (173)</td>
<td>19 (14)</td>
</tr>
<tr>
<td>3</td>
<td>4781 (344)</td>
<td>48 (25)</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>4</td>
<td>4186 (201)</td>
<td>23 (14)</td>
<td>106 (30)</td>
<td>433 (34)</td>
<td>3560 (173)</td>
<td>64 (16)</td>
</tr>
<tr>
<td>8</td>
<td>3518 (307)</td>
<td>10 (11)</td>
<td>0 (0)</td>
<td>61 (25)</td>
<td>3276 (287)</td>
<td>87 (11)</td>
</tr>
<tr>
<td>11</td>
<td>2427 (202)</td>
<td>0 (0)</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>15</td>
<td>1342 (185)</td>
<td>33 (13)</td>
<td>-</td>
<td>-</td>
<td>1162 (160)</td>
<td>63 (10)</td>
</tr>
<tr>
<td>18</td>
<td>1233 (302)</td>
<td>0 (0)</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>22</td>
<td>752 (53)</td>
<td>0 (0)</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>32</td>
<td>950 (234)</td>
<td>13 (16)</td>
<td>-</td>
<td>-</td>
<td>141 (51)</td>
<td>735 (194)</td>
</tr>
</tbody>
</table>

"-" : no available data. Values shown are the mean and standard error of the mean (sem) for estimated numbers of different types of cells.

we obtain


Therefore, we can estimate mean and variance of number of BrdU+ labeled cells of any type \( i \) \((E[X] \text{ and } V[X])\) by equations below:

\[
E[X] = \bar{Y} \times \bar{P}
\]
\[
V[X] = S_p^2(\bar{Y}^2 + S_Y^2 - \bar{Y}) + \bar{P}^2(S_P^2 - \bar{Y}) + \bar{Y} \times \bar{P}
\]

A summary of data in estimated cell numbers is shown in Table 4.5

4.4.3 Initial distribution of cell population

Prior to deriving the model to simulate labeling curves, we focus on the quantification of the cell population at the moment of the BrdU injection and aim to obtain the distribution of cell population at \( t = 0 \).
In matrix \( M \) (Eq. 4.1), the fundamental solution of the model \( M_{ij} \) as a function of time allows finding the number of cells at time \( t \) in compartment \( j \), given that the population was seeded by a single cell in compartment \( i \). However, under physiological conditions, the system is fed by a steady influx of freshly activated ANPs. Under such assumption, the number of each cell type at time \( t \) is given by

\[
\tilde{M}(t) = \lambda \sum_{k=0}^{\infty} [T(t)m]^k \ast \int_0^t [(I - T(\tau))]d\tau
\]  

(4.3)

Since the system is fed only by newborn ANPs at their first division in \( G_1 \)-phase, only the top row of matrix \( \tilde{M}(t) \), denoted by \( \tilde{M}^{(1)}(t) \), is required. In labeling experiments, we treat 1 month old mice with BrdU injections. We assume that the snapshot of the neurogenic cell population in the animal’s brain is under a “steady state” at the moment of injection. Thus, if we take \( t \) to infinity in the equation  (4.3), we can obtain a stationary distribution of cell numbers in the snapshot, which yields numbers of cells of different types, given by

\[
\pi = \tilde{M}^{(1)}(\infty) = \lim_{t \to \infty} \lambda \sum_{k=0}^{\infty} [T(t)m]^k \ast \int_0^t [(I - T(\tau))]d\tau \]  

(1)

\[ (1) \]  

where \( \pi \) is a vector of numbers of cells in all modeling compartments (see Matrix 4.2 for the description of modeling compartments), diagonal matrix \( E[T] \) has entries equal to expected duration times in all modeling compartments, \(-1\) is the notation of inverse of matrix, \((1)\) means the top row of the matrix and \( \lambda, T, m, I, \ast \) are as previously defined. Although in the long run, the intensity parameter \( \lambda \) will decline as the animal gets older, we can still assume that in a short period of time,
within which the snapshot of the BrdU injection occurs, the influx rate of QNPs to newborn ANPs is constant.

BrdU labels all cells that are in S-phase, thus we know how many cells are labeled at the moment of injection, which is equal to $\pi_s$, where subscript $s$ stands for the compartment or a set of compartments which represents cells in the S-phase. Under the assumption that in vivo descendants of labeled cells remain labeled (label dilution is neglected), the pulse labeling curve of number of labeled cells at given time is derived as $\pi_s M(t)$. This expression is technically true only under the assumption that labeled cells are grouped at the beginning of the S-phase, which does not cause much difference for times longer than the joint duration of $S$ and $G_2M$. In real population, the remaining time for each cell in S-phase at any moment is a random variable. Additionally, the $k$-fold convolution of matrix $(Tm)$ in matrix $M(t)$ (Eq. 4.1) is becoming analytically too complicated as $k$ increases. Therefore, instead of obtaining the time-course labeling curve analytically, we decide to generate it by simulation in a more convenient and straightforward manner.

### 4.4.4 Simulation of labeling curves and parameter search

For obtaining the set of model parameters that yields best fit to the experimental observation of BrdU pulse-and-chase labeling curves, we have applied the genetic algorithm as the search heuristic in the simulation program to explore a realistic parameter space that specifies a variety of discrete values within a range to each of the key model parameters (see Table 4.6 for a list of model parameters and their assigned ranges of possible values). For example, the maximum number of ANP divisions ranges from 2 to 8, renewal probability of ANP cell ranges from 0 to 1,
expected duration of ANP S-phase ranges from 5 to 12 hr, shape parameter of distribution of ANP S-phase duration ranges from 5 to 40, minimum duration of ANP S-phase ranges from 1 to 4 and apoptotic rate for ANP S-phase can range from 0 to 0.99.

### 4.4.4.1 Data fitting

Simulation results of BrdU labeling curves that can best fit our data are illustrated in Figure 4.5. Parameter values that yield such fit are obtained by the genetic algorithm, which is implemented for the purpose of searching the parameter space.

---

**Table 4.6: List of model parameters and ranges of possible values**

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Range of possible values</th>
</tr>
</thead>
<tbody>
<tr>
<td>Minimum number of ANP divisions, $min_{ANP}$</td>
<td>1, 2, 3</td>
</tr>
<tr>
<td>Maximum number of ANP divisions, $max_{ANP}$</td>
<td>2, ..., 8</td>
</tr>
<tr>
<td>Renewal probability of ANP, $p_{ANP}$</td>
<td>( {0, 0.1, ..., 0.99, 1} )</td>
</tr>
<tr>
<td>Distribution coefficients of ANP $G_1$-phase duration, $T_{G_1-ANP}$</td>
<td>( {6, ..., 20} \text{ hr}; {2, ..., 16}; {2, ..., 5} \text{ hr} )</td>
</tr>
<tr>
<td>Distribution coefficients of ANP S-phase duration, $T_{S-ANP}$</td>
<td>( {5, ..., 12} \text{ hr}; {5, ..., 40}; {1, ..., 4} \text{ hr} )</td>
</tr>
<tr>
<td>Distribution coefficients of ANP $G_2M$-phase duration, $T_{G_2M-ANP}$</td>
<td>( {1, ..., 4} \text{ hr}; {5, ..., 20}; {0, ..., 0.75} \text{ hr} )</td>
</tr>
<tr>
<td>Distribution coefficients of ANP-NB stage duration, $T_{ANP-NB}$</td>
<td>( {4, ..., 64} \text{ hr}; {2, ..., 16}; {0, ..., 3} \text{ hr} )</td>
</tr>
<tr>
<td>Distribution coefficients of ANP-Apop stage duration, $T_{ANP-Apop}$</td>
<td>( {4, ..., 64} \text{ hr}; {2, ..., 16}; {0, ..., 3} \text{ hr} )</td>
</tr>
<tr>
<td>Distribution coefficients of NB duration, $T_{NB}$</td>
<td>( {120, ..., 430} \text{ hr}; {2, ..., 16}; {10, ..., 80} \text{ hr} )</td>
</tr>
<tr>
<td>Distribution coefficients of Apoptotic cell duration, $T_{Apop}$</td>
<td>( {0.4, ..., 3} \text{ hr}; {2, ..., 16}; {0, ..., 0.3} \text{ hr} )</td>
</tr>
<tr>
<td>Cell death rate of ANP $G_1$-phase, $d_{G_1}$</td>
<td>( {0.0, 0.1, ..., 0.98, 0.99} )</td>
</tr>
<tr>
<td>Cell death rate of ANP S-phase, $d_S$</td>
<td>( {0.0, 0.1, ..., 0.98, 0.99} )</td>
</tr>
<tr>
<td>Cell death rate of ANPG2M-phase, $d_{G2M}$</td>
<td>( {0.0, 0.1, ..., 0.98, 0.99} )</td>
</tr>
<tr>
<td>Cell death rate of non-proliferating ANP, $d_{ANP}$</td>
<td>( {0.0, 0.1, ..., 0.98, 0.99} )</td>
</tr>
<tr>
<td>Cell death rate of NB, $d_{NB}$</td>
<td>( {0.0, 0.1, ..., 0.98, 0.99} )</td>
</tr>
<tr>
<td>Minimum number of QNP divisions, $min_{QNP}$</td>
<td>( {1, 2, 3} )</td>
</tr>
<tr>
<td>Maximum number of QNP divisions, $max_{QNP}$</td>
<td>( {2, ..., 6} )</td>
</tr>
<tr>
<td>Renewal probability of QNP, $p_{QNP}$</td>
<td>( {0, 0.1, ..., 0.99, 1} )</td>
</tr>
<tr>
<td>Distribution coefficients of QNP $G_1$-phase duration, $T_{G_1-QNP}$</td>
<td>( {8, ..., 36} \text{ hr}; {2, ..., 16}; {2, ..., 5} \text{ hr} )</td>
</tr>
<tr>
<td>Distribution coefficients of QNP S-phase duration, $T_{S-QNP}$</td>
<td>( {5, ..., 12} \text{ hr}; {5, ..., 40}; {1, ..., 4} \text{ hr} )</td>
</tr>
<tr>
<td>Distribution coefficients of QNP $G_2M$-phase duration, $T_{G2M-QNP}$</td>
<td>( {1, ..., 4} \text{ hr}; {5, ..., 20}; {0, ..., 0.75} \text{ hr} )</td>
</tr>
</tbody>
</table>

---

1If a non-proliferative ANP is determined to differentiate to a neuroblast it enters ANP-NB stage, otherwise it enters ANP-Apop stage before undergoing apoptosis. Since the cell duration (transit time) is modeled by a shifted gamma distribution the duration distribution parameter for any cell type $i$, $T_i$, consists of 3 coefficients that are expected duration, shape parameter of the gamma distribution and the minimum duration (shift value). A range of values for each of these three coefficients has been provided.
Simulation results (red solid lines) that can best fit data (black dashed lines) of all available measurements from running the genetic algorithm. “total-BrdU+” stands for total number of BrdU+ cells, “Apop-BrdU+” for number of BrdU labeled apoptotic cells, “QNP-BrdU+” for number of BrdU labeled QNP cells, “ANP-BrdU+” for number of BrdU+ ANP cells, “NB-BrdU+” and “GC-BrdU+” for numbers of BrdU+ NB and GC cells, respectively. On each plot the shaded area depicts the region that is upper and lower bounded by the average cell count plus or minus 2 sems.

and optimizing the goodness-of-fit function. BrdU pulse labeling experiments have provided us 40 non-trivial (non-zero) independently measured experimental data points. There are 19 model parameters that are varied during the parameter search. Figure 4.6 demonstrates that the residuals are equally distributed along $x$ axis and show no systematic trend, which indicates that the model fit is good.

4.4.4.2 Summary of apoptotic rates

Among all model parameters, apoptotic rates are most crucial ones that have not been analyzed or estimated in any early-stage hippocampal neurogenesis studies before. Figure 4.7 shows a comparison of estimated apoptotic rate by our model
Residuals plot that shows the distribution of differences (difference = data - simulated result) on all non-zero measurements.

and simulation results at each cell state through early stages of hippocampal neurogenesis in 1 month old mice brains. Apoptosis rates are low in proliferating ANP cells whereas once ANP become non-proliferating about one third of them will undergo apoptosis. During NB stage apoptosis reaches maximum. A vast majority of neuroblasts will die and only few of them (estimated about 3%) will differentiate into granule neurons (97% undergo apoptosis). QNP do not enter apoptosis (Sierra et al., 2010) and once a QNP cell is activated it undergoes a number of asymmetric divisions and eventually becomes an astrocyte.

4.4.4.3 Comparison of parameter values with literature

Encinas et al. (2011) also carried out labeling experiments (both single and double labeling) to study adult hippocampal neurogenesis. They modeled neurogenic cascade similarly as we do, although they used 2 month old mice whereas we
Comparison of the estimated apoptotic rate at each cell state through early stages of hippocampal neurogenesis in 1 month old mice brains. “np-ANP” stands for non-proliferating ANP. Cell death rates are low in amplifying progenitor cells (ANP) and can reach very high in neuroblasts (NB).

used 1 month old ones. The major difference is that they did not consider apoptosis. They determined division and duration related parameters, and provided estimated values, such as expected duration of each type of cell. Their results were calculated from inferring the decay rate of each type of cell over a long period of time (800 days). Parameter values that yield best fit in our study are listed in Table 4.7, where most of expected cell durations are comparable with Encinas et al.’s estimates. In addition, our model and simulation approach is also able to provide estimates on apoptotic rates, minimum durations, shapes of duration distributions, and numbers of QNP and ANP divisions.

4.4.5 Prediction of dynamics of neurogenesis under reduced apoptosis

Taking values of model parameters from data fitting results we carry out additional
Table 4.7: Parameter estimates that yield best fit and comparison with estimates in literature

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Value</th>
<th>Estimate of Encinas et al.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Minimum number of ANP divisions</td>
<td>1</td>
<td>-</td>
</tr>
<tr>
<td>Maximum number of ANP divisions</td>
<td>4</td>
<td>-</td>
</tr>
<tr>
<td>Renewal probability of ANP</td>
<td>0.1</td>
<td>-</td>
</tr>
<tr>
<td>Expected number of ANP divisions</td>
<td>1.17</td>
<td>2</td>
</tr>
<tr>
<td>Minimum ANP $G_1$-phase duration</td>
<td>12hr</td>
<td>-</td>
</tr>
<tr>
<td>Expected ANP S-phase duration</td>
<td>12hr</td>
<td>12hr</td>
</tr>
<tr>
<td>Minimum ANP S-phase duration</td>
<td>4hr</td>
<td>-</td>
</tr>
<tr>
<td>Expected ANP $G_2M$-phase duration</td>
<td>1hr</td>
<td>2hr</td>
</tr>
<tr>
<td>Minimum ANP $G_2M$-phase duration</td>
<td>0.5hr</td>
<td>-</td>
</tr>
<tr>
<td>Expected ANP-NB$^1$ duration</td>
<td>12hr</td>
<td>30hr</td>
</tr>
<tr>
<td>Minimum ANP-NB duration</td>
<td>3hr</td>
<td>-</td>
</tr>
<tr>
<td>Expected ANP-Apop$^2$ duration</td>
<td>48hr</td>
<td>-</td>
</tr>
<tr>
<td>Minimum ANP-Apop duration</td>
<td>2hr</td>
<td>-</td>
</tr>
<tr>
<td>Expected NB duration</td>
<td>260hr</td>
<td>60hr &amp; 306hr$^3$</td>
</tr>
<tr>
<td>Shape parameter of NB duration distribution</td>
<td>2</td>
<td>-</td>
</tr>
<tr>
<td>Minimum NB duration</td>
<td>20hr</td>
<td>-</td>
</tr>
<tr>
<td>Expected apoptotic cell duration</td>
<td>1.4hr</td>
<td>-</td>
</tr>
<tr>
<td>Cell death rate of ANP $G_1$-phase</td>
<td>0.14</td>
<td>-</td>
</tr>
<tr>
<td>Cell death rate of ANP S-phase</td>
<td>0</td>
<td>-</td>
</tr>
<tr>
<td>Cell death rate of ANP $G_2M$-phase</td>
<td>0.02</td>
<td>-</td>
</tr>
<tr>
<td>Cell death rate of nonproliferating ANP</td>
<td>0.33</td>
<td>-</td>
</tr>
<tr>
<td>Cell death rate of NB</td>
<td>0.97</td>
<td>-</td>
</tr>
<tr>
<td>Minimum number of QNP divisions</td>
<td>2</td>
<td>-</td>
</tr>
<tr>
<td>Maximum number of QNP divisions</td>
<td>5</td>
<td>-</td>
</tr>
<tr>
<td>Renewal probability of QNP</td>
<td>0.57</td>
<td>-</td>
</tr>
<tr>
<td>Expected number of QNP divisions</td>
<td>3.57</td>
<td>3</td>
</tr>
<tr>
<td>Expected QNP $G_1$-phase duration</td>
<td>28hr</td>
<td>-</td>
</tr>
<tr>
<td>Expected QNP S-phase duration</td>
<td>11hr</td>
<td>8hr</td>
</tr>
<tr>
<td>Expected QNP $G_2M$-phase duration</td>
<td>3hr</td>
<td>2hr</td>
</tr>
<tr>
<td>Expected QNP duration</td>
<td>42hr</td>
<td>28hr, 28hr and 52hr$^4$</td>
</tr>
</tbody>
</table>

$^1$ANP-NB is the transition stage between ANP and NB.
$^2$ANP-Apop is the transition stage between ANP and apoptotic cells.
$^3$NB durations for $t < 100hr$ and $t > 100hr$, respectively.
$^4$Expected durations of first, second and third divisions.
simulations to predict the overall changes in BrdU labeling curves by inhibiting apoptosis (reducing apoptotic rates). While apoptotic rates at all cell stages are consistently reduced by a hypothetical amount (25%, 50%, 75% or 100%) all other model parameters remain unchanged.

From predicted BrdU labeling curves depicted in Figure 4.8 we observe that at the end of 32 days total number of BrdU+ cells and number of BrdU labeled granule cells increase 3.4 and 11.5 folds, respectively, when apoptotic rates are reduced by 25% only. These numbers continue to increase sharply if apoptosis can be reduced even further. Under the extreme situation when apoptosis can be completely inhibited the simulation results indicate that 14.3 times more of total BrdU+ cells and 61.0 times more of BrdU+ GCs are expected compared with the case where normal apoptotic rates are employed. Our study indicates that reducing apoptosis significantly increases adult hippocampal neurogenesis.
Figure 4.8: Prediction of dynamics of neurogenesis under reduced apoptosis

Effect of reduced apoptosis on simulated labeling curves of different types of cells over the time course of 32 days (black - normal apoptotic rates; red - apoptotic rates reduced to 75% of normal rates; green - apoptotic rates reduced to 50% of normal; blue - apoptosis reduced to 25%; light blue - apoptosis completely inhibited to 0%).
Chapter 5

Modeling of stochastic events related to DNA content distributions in proliferating cells

5.1 Motivation and overview

In this chapter, we demonstrate that experimental asymmetry in DNA-replication scatterplots is the hallmark of an increasing replication initiation rate in the S-phase of cell cycle. Understanding kinetics of DNA replication gives an insight into mechanisms revealing specifics of normal and cancer cells proliferation. Mathematical modeling allows relating molecular events to single-cell characteristics assessed by multiparameter cytometry. In the present study we labeled newly synthesized DNA in A549 human lung carcinoma cells with 15 – 120 min pulses of EdU. All DNA was stained with DAPI and cellular fluorescence was measured by laser scanning cytometry. The “horseshoe”-shaped bivariate scatterplots of EdU incorporation vs. DAPI level report the rate of DNA synthesis and the DNA content, respectively. To understand the connection between molecular-scale events and scatterplot asymmetry, we develop a multiscale stochastic model, which simulates DNA replication and cell cycle progression of individual cells and produces \textit{in silico} EdU/DAPI scatterplots. For each S-phase cell the time points at which replication origins are
fired are modeled by a non-homogeneous Poisson Process (NHPP). Shifted gamma distributions are assumed for durations of cell cycle phases ($G_1$, $S$ and $G_{2M}$). Depending on the rate of DNA synthesis being an increasing or decreasing function, simulated EdU/DAPI graphs show predominance of cells in left (early-S) or right (late-S) side of the horseshoe distribution. If the rate of DNA synthesis is constant, the EdU/DAPI graphs are symmetric. Assuming NHPP rate estimated from independent experiments, simulated EdU/DAPI graphs are nearly indistinguishable from those experimentally observed. This finding proves consistency between the S-phase DNA-replication rate based on molecular-scale analyses, and cell population kinetics ascertained from EdU/DAPI scatterplots. Our approach opens a possibility of similar modeling to study the effect of anticancer drugs on DNA replication/cell cycle progression and also to quantify other kinetic events that can be measured during S-phase.

5.2 Introduction

High-throughput techniques such as DNA combing and massively parallel sequencing, have generated vast amounts of data concerning spatiotemporal patterns of DNA replication in a number of organisms (Hyrien and Goldar, 2010) [59]. These data, based on single-cell analysis and population-profiling, allow building comprehensively mathematical models of DNA replication process. Recent studies focused on a large number of quantities which are of biological interest, such as whole-genome replication timing profiles, replication fork velocity, average number of active origins, origin activation rate, replicon sizes, and other (Retkute et al., 2012)
Among the most challenging problem is understanding the mechanisms that coordinate the temporal order of activation of replication origins. Published results favored either stochastic models of replicons fired independently of each other or correlated-origin models with adjacent origins fired at similar times (Guilbaud et al., 2011). Although the temporal order of activation of origins is never exactly the same in any two cells, it has been found that the general pattern of the replication process is preserved, that is the rate of origin firing has a universal shape which increases during early to mid-S-phase and decreases at mid to late-S-phase (Ma et al., 2012). In our study, we use stochastic modeling to determine if single-cell measurements of the rate of DNA replication are confirmed in cycling cell populations.

Flow cytometry, LSC, and other earlier methods have been long used to measure the rate of DNA replication across duration of the S-phase of the cell cycle. Early studies made use of the DNA fiber autoradiography method. Specifically, the Chinese hamster ovary cells arrested (synchronized) at the entry to S-phase had been released from the arrest, and at different time intervals after the release were pulsed with tritiated thymidine; [3H]dT (Housman and Huberman, 1975). It was concluded that the rate of replication (fork movement) was approximately three times lower during the first hour than during the final hour of the S-phase. Using similar approach the rate of progression of DNA replication forks was estimated in normal epidermal cells and compared to basal cell carcinoma (Galand and Heenen, 1988); however, variability of the progression across the S-phase was not estimated. Another approach to explore DNA replication rate during S-phase was based on cell
labeling with [3H]dT followed by cell sorting and autoradiography, which has shown that the early S-phase cells had approximately two times lower DNA replication rate compared to mid or late S-phase cells (Clausen, 1987) [23]. Still another approach to measure DNA replication rate was based on analysis of replication of the 400 copies (43 kb) of a ribosomal DNA located on five acrocentric chromosomes 13, 14, 15, 21 and 22 replication of which spans most of the S-phase duration (Scott et al., 1997) [106]; this suggested that the rate of replication of rDNA was also lower at the early- compared to the late S-phase.

Immunocytochemical detection of 5-bromo-2’-deoxyuridine (BrdU) incorporation combined with the DNA content measurement and analysis of DNA content/BrdU incorporation bivariate distributions opened new possibilities to measure kinetics of the cell cycle progression (Dolbeare et al., 1983) [36]. Pulse labeling with BrdU followed by different time intervals of cell growth in the absence of BrdU (“pulse-chase labeling”) made it possible to measure, with high accuracy, the “relative movement” of the pulse-labeled cells across the cell cycle (Begg et al., 1985) [9]. An attempt has been made to use this approach to estimate variability in DNA replication rates across the S-phase (Bertuzzi et al., 1995) [10]; however without experimental confirmation.

There are two potential difficulties shared by the previous estimates of the variability of DNA replication rate across the S-phase. In some of the studies (Housman and Huberman, 1975; Galand and Heenen, 1988; Scott et al., 1997) [57, 47, 106] cells were synchronized at the entrance to the S-phase by inhibitors of DNA polymerase. This type of synchronization induces DNA replication stress manifesting
itself by activation of DNA damage signaling (Kurose et al., 2006) [73] and also induces unbalanced growth and unscheduled expression of cyclins B1, E, A and D3 (Gong et al., 1995) [50]. Since such changes may affect DNA replication rate, extrapolating data from so treated cells may not be representative of the untreated and unperturbed exponentially growing cells. Another problem stems from the use of BrdU as DNA precursor. Immunocytochemical detection of the incorporated BrdU requires partial denaturation of DNA, to make BrdU on the denatured DNA sections accessible to BrdU Ab antibody. At the same time a part of DNA has to remain non-denatured to be stainable with the intercalating dyes such as propidium iodide (PI) or 7-aminoactinomycin D (AAD). Because in situ DNA susceptibility to denaturation by acid or heat varies depending on chromatin structure (Darzynkiewicz, 1994) [27] the relative proportions of the denatured (binding BrdU Ab) versus non-denatured (stainable with PI) DNA may vary. Furthermore, since BrdU Ab is a rather large molecule, its accessibility to the incorporated BrdU is further restricted by steric hindrance. As a result, the immunocytochemical detection of the incorporated BrdU may not be stoichiometrically related to BrdU incorporation and thus to DNA replication rate. The use of [3]HdT in the early studies may have also affected DNA replication due to possible DNA damage by the disintegrating radioactive precursor.

Alternative more recent method to estimate DNA replication, based on the use of 5-ethynyl-2’deoxyuridine (EdU) as a DNA precursor (Salic and Mitchison, 2008) [105], offers several advantages over the BrdU incorporation approach (Darzynkiewicz et al., 2011) [29]. This precursor, once incorporated into DNA, is detected with
fluorochrome-tagged azides by means of a copper (I) catalyzed [3+2] cycloaddition reaction defined as “click chemistry” (Meldal and Tornøe, 2008) [81]. Small size of the azides that much more easily penetrate into the specimen than BrdU antibodies offers one advantage, viz. greater accessibility of the incorporated EdU to the fluorochrome. Another advantage stems from the fact that DNA denaturation is not required. Therefore DNA content is measured stoichiometrically. Detection of EdU by this methodology is also expected to be much more stoichiometric vis-à-vis the actual extent of EdU incorporation (DNA replication rate) than the BrdU incorporation is. In the present study we applied this methodology combined with laser scanning cytometry (LSC) (Pozarowski et al., 2013) [94] to assess the rate of cell progression through the early and late sections of S-phase.

We address the experimental cytometric data based on EdU incorporation, by multiscale mathematical modeling of variability of incorporation of the precursor across the span of the S-phase. This methodology produces novel insights into the kinetics of DNA replication. Among a number of stochastic and semi-stochastic approaches that have been used to model cell cycle progression, a few use data derived from cytometry labeling experiments (Basse et al., 2005; Larsson et al., 2008; Bertuzzi et al., 2002) [7, 74, 11] (also see Discussion). These models provide estimates of DNA replication rates based on analyzing and evaluating BrdU-DNA flow cytometric histograms in the pulse-labeling cell sample scheme.

In our mathematical model, we ensure that each locus is replicated once and only once per cell cycle. In our modeling scheme of DNA replication per cell, its entire genome is evenly divided into N DNA segments, each segment has $M = L/N$.
nucleotide base pairs, where \( N \) is the number of licensed (or selected) replication origins/replicons (Hyrien et al., 2003; Jun et al., 2004) and \( L \) is the length of the whole genome. Each replicon is fired (activated) at a different time point, which is generated conditionally on the duration of the S-phase. This “mathematical trick” allows first to determine the duration of S and then to “fit” all needed replication events in S. Once a replicon has been fired, it replicates \( M \) base pairs of DNA. Thus, each locus is replicated once and only once per cell cycle.

5.3 Methods

5.3.1 Laboratory procedures

Human lung carcinoma A549 cells were purchased from American Type Culture Collection (ATCC CCL-185, Manassas, VA). The cells were cultured in Ham’s F12K medium with 2mM L-glutamine, adjusted to contain 1.5g/L sodium bicarbonate (ATCC) and supplemented with 10% fetal bovine serum (ATCC). Dual-chambered slides (Nunc Lab-Tek II) were seeded with 1 mL of 105 cells/ml in each chamber 24h (to ~50% confluency) prior to exposure to EdU. All incubations were carried out at 37°C in a humidified atmosphere of 5% CO\(_2\) in air. The cultures were treated with 10\( \mu \)M EdU (Invitrogen/Molecular Probes, Eugene, OR) for periods of time as shown in Figure 5.1. After exposure to EdU the cells on slides were fixed by transferring slides into Coplin jars containing 1% methanol- free formaldehyde (Polysciences, Warrington, PA) in PBS for 15min on ice, then rinsed with PBS and stored in 1% BSA in PBS, at 4°C for up to 24h. The Click-iTTM EdU Alexa Fluor 488 imaging kit (Invitrogen/Molecular Probes) was used to detect EdU incorpo-
ration. Before measurement by LSC, the cells were counterstained with 2.8 µg/ml 4,6-diamidino-2-phenylindole (DAPI; Sigma Chemical Co, St Louis MO) in PBS for 15 min. Cellular green (Alexa Fluor 488) and blue (DAPI) emission was measured using an LSC (iCys; CompuCyte, Westwood, MA) utilizing standard filter settings; fluorescence was excited with 488-nm argon, helium-neon (633 nm) and violet (405 nm) lasers (Pozarowski et al., 2013). The intensities of maximal pixel and integrated fluorescence were measured and recorded for each cell. At least 3,000 cells were measured per sample. Further details are given in Figure 5.1 legend.

5.3.2 Computations

5.3.2.1 Transformation of LSC data

Raw data have the form of a set of numerical coordinates of individual cells, the first of which \((x)\) is the DAPI level and the second \((y)\) is the EdU incorporation level. Units are arbitrary and depend on the manual adjustment of instrument’s gain. In the raw data, the right arm of the “horseshoe” graph is elevated with respect to the left arm, one of the reasons being non-specific staining. This effect is distracting for visual comparisons with simulated data. Therefore, we use an ad hoc linear transformation \(y' = \alpha y + \beta\) with coefficients \(\alpha\) and \(\beta\) and adjusted so that the centroids of the \(G_1\) and \(G_{2M}\) clusters have identical coordinates (corresponding to no EdU incorporation). The resulting scatterplots are depicted in Figure 5.1. The vertical coordinate is further scaled so that the centroid of the upper part of the horseshoe is approximately at the same level at each panel in Figure 5.1.
Bivariate DNA content versus EdU incorporation distributions of A549 cells incubated with EdU for 15, 30, 45, 60 and 120 min. Based on differences in DNA content and EdU incorporation cells in G\textsubscript{1}, S and G\textsubscript{2}M can be identified. Furthermore, the cells with variable level of EdU incorporation and with DNA content close to that of G\textsubscript{1} (eS) during the EdU pulse were entering S-phase. The cells with variable level of EdU incorporation and DNA content close to that of G\textsubscript{2}M (eG\textsubscript{2}M) cells were exiting S and entering G\textsubscript{2}M during the pulse. The cells with maximal EdU level (S) were exposed to EdU for the full duration of the pulse. Note that regardless of the EdU pulse duration, the frequency of eS cells markedly exceeds frequency of eG\textsubscript{2}M cells. This indicates that the rate of cells progression through very early segment of S-phase must be slower compared with the rate of their progression through late S-phase segment.
5.3.2.2 Modeling and simulation steps

Modeling consists of two steps. The first step is to perform simulation of *in vitro* cell growth to mimic continuous cell culture, periodically passaged to reach the ‘dynamic equilibrium’ of exponentially growing cell population, in which proportions of cells in different cell cycle phases remain the same, and from which cell samples are further drawn. The second step is to simulate DNA labeling. For each dividing cell captured by EdU during the S-phase, the time points at which replicons are fired are modeled by a non-homogeneous Poisson Process (Ross, 1996) [101].

5.3.2.3 Cell transit times

Cell transit time is defined as the time that a cell spends in a given cell cycle phase or stage before it proceeds to the next phase or stage. Instead of the commonly used exponential distributions, we employ the shifted gamma distributions of the transit times of the form

\[
f(t|k, s, v) = (t-v)^{k-1} \exp\left[-(t-v)/s\right]/\left[s^k \Gamma(k)\right], \quad (t > v)
\]

(Casella and Berger, 2001) [18] to model the cell transit times through phases G₁, S and G₂M. Advantage of these latter is the ability to independently specify the minimum transit times, mean transit times, and variances of transit times. Among other, this allows avoiding occurrence of cells that live for an indefinitely short period of time with certain probability. Equivalently, there are three free parameters; is the shape parameter, is the scale parameter and is the shift value (transit time); is Euler gamma function (Casella and Berger, 2001) [18]. Cells in different cell cycle phases are assigned shifted gamma distributions with different sets of parameters.
5.3.2.4 Modeling and simulation of in vitro cell culture

As a preparatory step, a cell population is created in silico as a set of proliferating cells that have been synchronized at the beginning of their respective G_1-phases. Each cell is identified by three parameters in order to uniquely record its status at any given time point (discretized to one-minute units). These are ‘phase’, ‘time spent’ and ‘time remained’ that indicate in which cycling phase the cell currently resides, how much time it has already spent in that phase and how much time is required for the cell to transit into the next phase, respectively. Subsequently, the simulated cell population is evolved cycling for a long period of time with cells proliferating, until it achieves exponential growth and a ‘dynamic equilibrium’ in which the proportion of each type of cells remains unchanged. The simulation program tracks the behavior of each cell every time unit by updating its parameter values and removes a randomly selected fraction of cells from the population whenever the population size reaches a set limit so as to mimic cell passaging procedures. Key underlying assumptions are as follows: cells proceed through the G_1, S and G_2M-phases; duration of each phase is sampled from a shifted gamma distribution. Cell death is not considered in the current model, but it can be readily considered. At the end of this preparatory period a snapshot of the cell population is saved from which cell samples are drawn as the initial (time = 0) samples for the labeling experiment.

5.3.2.5 Modeling DNA replication

The important mathematical tool we use to model the process of DNA replication is the non-homogeneous Poisson process (NHPP). Intuitively, the NHPP is the
sequence of time events appearing at random intervals, with intensity that itself is a function of time (hence the term “nonhomogeneous”). NHPP is particularly suitable since it can be defined conditionally on a given number of events (such as the number of replicons $N$) restricted to a given time interval (duration of $S$) and at the same time be endowed with a rate following a prescribed time pattern scaled to S-phase duration. In this way, it is ensured that all the replication origins “fit into” the phenomenologically defined S-phase.

More specifically, as explained earlier on, for each cell we draw a random number from a predetermined shifted-gamma distribution as the transit time through S-phase, $T_s$, and a random number from a predetermined normal distribution as the fork velocity, $\nu$. The cell’s initiation rate of replicon firing is following the shape of the chosen NHPP rate function, under the constraints that (i) the rate function is defined on the interval $[t, t + T_s - M/\nu]$, where $t$ is the chronological time of beginning of the S-phase (end time of the previous G$_1$-phase), and $M = L/N$, where $L$ is the length of the whole genome and $N$ is the number of replicons, (ii) the first activated replicon is fired at $t$, (iii) once fired, each replicon replicates equal amount of DNA, $M$, with velocity $\nu$, and (iv) the last activated replicon is fired at $t + T_s - M/\nu$. Accordingly, the origination time of each replicon generated by NHPP is contained within the range of $[t, t + T_s - M/\nu]$ to preserve the pattern of firing rate of replicons and to avoid the “random completion” problem (Hyrien et al., 2003) [60].

Technically, origination times of replicons (DNA synthesis origins) are modeled by a non-homogeneous Poisson Process with a specified rate function $\mu(t)$ (Ross, 1996)
This means, among other things, that the expected number of replication origins becoming active (firing) in the time interval \((t, t + \Delta t)\) is equal to

\[
\int_t^{t+\Delta t} \mu(u)du \approx \mu(t)\Delta t + o(\Delta t)
\]

where \(o(\Delta t)\) becomes negligible faster than for \(\Delta t\) small. We assume here that each replicon replicates a fragment of DNA (length expressed in base pair units) at a constant speed (expressed in the units of base pair/second). Lengths and speeds are generated from normal distributions with pre-determined expectations and 10% coefficients of variation. The S-phase of a given cell starts from the initiation of its first replicon and ends with completion of replication of its whole genome. Alternative models of this process can be conceived (see Discussion) but they do not seem to add much in our context.

If the time-dependent rate of the Poisson Process is \(\mu(t)\), then the NHPP can be simulated using a transformation of the points of the time homogeneous Poisson Process with intensity 1 (HPP(1)), which are simulated in a standard manner. Briefly, if the points of the HPP(1) are \(t_0, t_1, t_2, \ldots\), then those of the NHPP are \(s_0, s_1, s_2, \ldots\), where \(s_i = M^{-1}(t_i)\) and \(M(\cdot)\) is the inverse function of the cumulative rate \(M(s) = \int_0^s \mu(u)du\). Accordingly, the relative overall ordering of firing of replicons for any cell follows a generally preserved pattern shaped by the choice of NHPP rate function, whereas the exact timing of each replicon differs among cells due to stochasticity of S-phase duration and the actual events generated by NHPP.
5.3.2.6 Modeling and simulation of cell labeling experiments

The aim is to generate in silico EdU/DAPI labeling levels acquired by all individual cells over the course of exposure. Given the exposure/labeling time $T = 15, 30, 45, 60$ or $120$ min, we draw a cell sample from the simulated ‘steady-state’ cell population and set time $T$. Immediately after the sample has been drawn (at $t = 0$) each cell is labeled by DAPI and EdU and resumes cycling. Subsequently, the simulation tracks and updates the labeling intensity and behavior of each cell every minute until the final simulation time $T$, at which information contained in the simulated cell sample is reported. It is assumed that there is no effect of labeling on cell cycle transit time or DNA synthesis rate. At fixation, a cell is categorized as ‘G\textsubscript{1}’ if it is in G\textsubscript{1}-phase at $t = T$, ‘eS’ (entering S) if in G\textsubscript{1} at $t = 0$ and in S at $t = T$, ‘S’ if in S at both $t = 0$ and $t = T$, ‘eG\textsubscript{2}M’ (entering G\textsubscript{2}M) if in S at $t = 0$ and in G\textsubscript{2}M at $t = T$, and as ‘G\textsubscript{2}M’ otherwise.

5.3.3 Simulation program

An event-based stochastic simulation program carrying out tasks as outlined above, has been written in Python programming language (http://www.python.org) with graphical user interface, and is available upon request. This program can be used to simulate both in vitro cell population and the labeling experiments. See the graphical user interface of the simulation program in Figure 5.2. Upon completion of simulations of the labeling experiments it generates scatterplots of EdU/DAPI bivariate distributions and DNA content distributions (DAPI staining intensity). Input parameters can be varied, such as duration distribution coefficients for G\textsubscript{1}, S
Python program to perform multiscale stochastic simulation which includes generating synchronized cell population, modeling DNA replication and simulating cell labeling experiments.

and G2M-phases; DNA replication rate (NHPP) function; total DNA length (base pairs) to be replicated in each cell; expected total number of replicons in each cell; expected replication speed of any replicon in each cell; labeling times, and other.

5.4 Results

5.4.1 Summary of LSC observations

Using the “click chemistry” approach we observe that following relatively short or longer (15 – 120 min) exposures to EdU, the bivariate distributions representing the amount of the incorporated EdU versus cellular DNA content (the integrated intensity of DAPI fluorescence) having a characteristic “horseshoe” pattern display with the right vs. left asymmetry with the left arm of the scatterplot being denser

Figure 5.2: Graphical user interface of simulation program for DNA replication and labeling experiments
Specifically, the number of cells with variable EdU content and DNA content close to that of G\textsubscript{1} cells (“horseshoe” left arm) is distinctly higher than that of cells with variable EdU content and DNA content close to that of G\textsubscript{2}M cells (“horseshoe” right arm) (Zhao et al., 2011). The left arm consists of cells that during duration of the exposure to EdU were entering S-phase (eS cells) and thus had variable duration of exposure to EdU at the time when they actually were replicating DNA. The right arm consists of cells exiting S and entering G\textsubscript{2} during the EdU pulse (eG\textsubscript{2}).

It should be noted that EdU enters cells and becomes incorporated rapidly even when a very short pulse of 10 seconds is applied (Fig. 5.10; see further on). Thus, the asymmetry in EdU incorporation reflected differences in the average cellular rates of DNA replication while entering or leaving the S-phase in the presence of EdU in culture. Based on this observation, we hypothesize that in the cycling cells in our experiment, the rate of progression (defined as the number of bases incorporated into nascent DNA per cell) through the early portion of the S-phase is slower compared with the rate of progression through the late portion of the S-phase (see Figure 5.1). Consistently, it is expected that with the same number of cells moving through the S-phase the probability of detecting cell that move through the “slow” section of the S-phase (eS) is greater than the probability of detecting cells that move through the “fast” section (eG\textsubscript{2}). It should be stressed that a similar type of asymmetry in the “horseshoe” pattern distributions was also detected previously in different types/lines of mammalian cells based on pulse labeling of exponentially growing cells with BrdU or iododeoxyuridine (IdU) as well as analysis by flow
cytometry (Darzynkiewicz et al., 2011; Zhao et al., 2011) [29, 127]. Therefore the “asymmetric” pattern as presently seen is neither unique to the cell type, use of EdU as the DNA precursor, nor to analysis by the imaging cytometry (LSC), although as mentioned earlier, the use of EdU offers less bias in reporting DNA replication rate compared to methodologies that utilize antibodies to detect the incorporated precursor.

5.4.2 Exploratory simulations

We first performed a set of simulations with different variants of the NHPP stochastic mechanism with the rates being linear functions of time in the S-phase, with different slope coefficients. A summary of parameter settings is presented in Tables 5.1 and 5.2.

Table 5.1: Input parameters for simulating in vitro cell culture

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of cells at the beginning</td>
<td>20,000</td>
</tr>
<tr>
<td>Allowed population size limit</td>
<td>100,000</td>
</tr>
<tr>
<td>Reset population size</td>
<td>50,000</td>
</tr>
<tr>
<td>Simulation time</td>
<td>1,000 hr</td>
</tr>
<tr>
<td>$G_1$ - phase duration distribution coefficients</td>
<td>$(k, s, v) = (8 hr, 2, 4 hr)$</td>
</tr>
<tr>
<td>$S$ - phase duration distribution coefficients</td>
<td>$(k, s, v) = (8 hr, 4, 6 hr)$</td>
</tr>
<tr>
<td>$G_{2M}$ - phase duration distribution coefficients</td>
<td>$(k, s, v) = (4 hr, 4, 3 hr)$</td>
</tr>
</tbody>
</table>

Table 5.2: Input parameters for simulating labeling experiments of DNA contents

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cell sample size</td>
<td>1,000</td>
</tr>
<tr>
<td>Total DNA length</td>
<td>$3 \times 10^9 bp$</td>
</tr>
<tr>
<td>Expected total number of replicons</td>
<td>30,000</td>
</tr>
<tr>
<td>Expected replication speed</td>
<td>100 bps/sec</td>
</tr>
<tr>
<td>EdU time of exposure</td>
<td>15, 30, 45, 60, 120 min</td>
</tr>
<tr>
<td>NHPP rate function</td>
<td>linear; increasing or decreasing</td>
</tr>
</tbody>
</table>

Concerning the specific procedure used to simulate NHPPs with linearly varying rates, we provide the example of the intensity function assumed to be linear (see
For each cell, $T$ denotes its S-phase duration, which is generated by a shifted gamma distribution. The average intensity is $R_{ave} = N/T$, where $N$ is the total number of replicons and $T$ is the S-phase duration. The ratio of minimum rate to average rate $r = R_{min}/R_{ave}$ (red – linearly increasing rates with $r = 1/2, 1/5$ and $1/10$; blue – linearly decreasing rates with $r = 1/2, 1/5$ and $1/10$; black – constant rate with $r = 1$).

Figure 5.3. We specify the total number of replicons $N$ and S-phase duration $T$, the average intensity $R_{ave} = N/T$. Given the ratio of minimum rate to average rate $r = R_{min}/R_{max}$, it is required that an additional condition $R_{max} = 2R_{ave} - R_{min}$ holds, in order to keep the integral of rate function constant on $[0, T]$ while $r \in (0, 1]$ is varied. We derive the linearly decreasing rate function

$$\mu(t) = 2N(r-1)t/T^2 + N(2-r)/T$$

where $t \in [0, T]$ and

$$M(s) = \int_0^s \mu(u)du = N(2-r)s/T + (r-1)Ns^2/T^2$$

If $r = 1$, $\mu(t)$ is a homogeneous Poisson Process.

Therefore, given the $i$th event $t_i$ of the time homogeneous Poisson Process $(0 < t_i \leq$
Simulation results assuming a constant rate function (black – $G_1$, red – $eS$, green – $S$, dark blue – $eG_2M$, light blue – $G_2M$). Constant intensity ($r = 1$).

$N_i$, we generate the $i$th event of the NHPP, $s_i$, by finding the root of $\frac{N(r-1)}{T^2} s_i^2 + \frac{N(2-r)}{T} s_i - t_i = 0$ if the rate is a linearly decreasing function.

Similarly, for the linearly increasing rate function, we can obtain $s_i$ from $\frac{N(1-r)}{T} s_i^2 + \frac{rN}{T} s_i - t_i = 0$.

Simulated EdU/DAPI scatterplots with $r = 1, 1/2, 1/5$ and $1/10$ for both cases of linearly decreasing and increasing rate functions are shown in Figures 5.4, 5 and 6. It is apparent from the inspection of the scatterplots that they are left asymmetric when a decreasing rate function is assumed, left vs. right asymmetric when an increasing rate is assumed and approximately symmetric when $r = 1$.

5.4.3 Simulations based on independently estimated intensity rate of the DNA synthesis NHPP

Encouraged by the outcome of the exploratory simulations, we carried out simulations using the NHPP intensity rate depicted in Figure 5.7. This figure was
Figure 5.5: Comparison of simulation results taking linear decreasing rate functions with different rates: (A) linear decreasing ($r = 1/2$), (B) linear decreasing ($r = 1/5$), and (C) linear decreasing ($r = 1/10$) (black – $G_1$, red – eS, green – S, dark blue – eG$_2$M, light blue – G$_2$M).
Comparison of simulation results assuming linear increasing rate functions with different rates: (A) linear increasing \( (r = 1/2) \), (B) linear increasing \( (r = 1/5) \), and (C) linear increasing \( (r = 1/10) \) (black – G\(_1\), red – eS, green – S, dark blue – eG\(_2\)M, light blue – G\(_2\)M).
Replication activity profile of a diploid human embryonic stem cell as derived from replication timing map (Desprat et al., 2009) [33], using methods as in Shaw et al., [107]. Analysis provided courtesy of Prof. Dean Jackson and Dr. Alex Shaw, University of Manchester. The DNA replication activity profile was created through fractioning the replication timing data set according to replication time into minute bins. The replication profile coordinates are obtained by measuring the amount of the different DNA sequences along the length of the chromosome in the increasing time order and plotted in a graphical form (see details in Desprat et al. [33]).

obtained from Prof. Dean Jackson and Dr. Alex Shaw of the University of Manchester, who used the TimEX data generated by Desprat et al. [33], processed as described by Shaw et al. [107], with genome wide replication expressed as a rate rather than overall amount of replication completed. Experimental methods are as in Refs. [33, 107]. As seen in the graph, the rate of this NHPP is bimodal, but with an overall tendency to increase. The scatterplots resulting are left vs. right asymmetric and visually indistinguishable from the experimental ones (Fig. 5.8 vs. Fig. 5.1).
Simulation results of labeling experiments via replication rate function derived from the replication activity profile (shown in Figure 5.7), following data of Ref. [33], constructed using methods of Ref. [107].

5.4.4 Analysis of DNA content frequency histogram

Analysis of the DNA content frequency histogram demonstrates an apparently uniform distribution of S-phase cells vis-à-vis their DNA content (DAPI staining intensity; Figure 5.9A). This is also evident when DNA content is measured with high accuracy, as provided by certain methods (Vindeløv and Christensen, 1994) [114]. It may be argued therefore that this is inconsistent with the rates of DNA replication being different for early and late sections of S-phase, as this latter effect would be reflected by DNA content histograms as an elevation at the early and depression at the late section of S-phase, or as a “hump” on the right side of the G\textsubscript{1} peak. The apparent lack of a hump on the right side of the G\textsubscript{1} peak or of distinct elevation of the early-S portion at DNA content histogram might be considered as an indication that in proportion to total cellular DNA the amount of DNA replicated at the slow rate is relatively small and therefore cannot account for a significant DNA content.
increase of these cells that would be evident on DNA histograms.

However, simulation modeling demonstrates that this observation is a reflection of lower sensitivity of single-variable DAPI histograms as compared to the bivariate DAPI/EdU scatterplots in terms of detection of cell cycle patterns. Indeed, as depicted in Figure 5.9C and 5.9D, simulated DAPI histograms corresponding to very different patterns of DNA synthesis, $r = 1$ (constant rate) and $r = 1/10$ (sharp linear increase in the rate), differ only slightly. At the same time, the corresponding DAPI/EdU scatterplots are very easy to distinguish by eye. The reason is the bivariate scatterplots’ higher information contents. Data simulated using DNA replication pattern (Figure 5.9B) based on reference Desprat et al., 2009 [33], do not show a visible hump in DAPI histogram, although they do show a visible asymmetry in the horseshoe (DAPI/EdU) scatterplot.

5.5 Discussion

The original feature of our study is that it confirms that the information obtained in single-cell measurements of the rate of DNA replication is consistent with the behavior of cycling cell populations, with mathematical modeling providing the connection. Our finding seems even more significant because it was obtained using A549 human non-small cell pulmonary adenocarcinoma cell line. This cell line has wild-type p53, represents a common type of malignant lung cancer and is widely used, particularly in studies of the effectiveness of anticancer drugs.

Over recent years, mathematical modeling has been used as an approach to study the principles of DNA replication in eukaryotes (Bechhoefer and Rhind, 2012) [8].
Comparison of DNA content distributions over different exposure times between experimental data and simulation results: (A) experiment, (B) simulation with replication rate function derived from Figure 5.7, (C) simulation with constant replication rate function, and (D) simulation with linear increasing rate function ($r = 1/10$).
Although the replication dynamics, which regulate replicon firing, is difficult to estimate experimentally it can be integrated into models at single-cell level and inferred by fitting the experimentally observable DNA replication timing patterns (Desprat et al., 2009) [33] (Figure 5.7). A widely accepted parsimonious hypothesis is that firing of each replicon is a stochastic event independent of the states of other replicons. Although a modeling scenario with origins being correlated has been proposed (Guilbaud et al., 2011; Shaw et al., 2010) [53, 107] based on the observation of clustering of replicons (Pasero et al., 2002) [87], all models are stochastic at the level of molecular interactions and recent experiments show that correlation between replicons has little impact on the general replication timing patterns (Bechhoefer and Rhind, 2012) [8]. A plausible conclusion from these models is that DNA synthesis rate is slow at the beginning, increases through S-phase duration and may become slow again when approaching the completion of replication.

As mentioned in the Introduction, several approaches have been used to estimate the rates of DNA replication at different sections of S-phase. In most of the early studies DNA replication rate was seen to be distinctly slower at the early-compared to the late-section of S-phase. Our present findings are consistent with these prior estimates. Cells synchronized by inhibitors of DNA replication were used in several prior studies. Because such synchronization induces replication stress (Darzynkiewicz et al., 2011) [28] and growth imbalance (Gong et al., 1995) [50], this approach might potentially introduce experimental bias. However, it appears that synchronization did not significantly affect the differences in the rate of DNA replication at the time of entrance to, versus exit from, the S-phase.
If a fully reductionist view is adopted, then the input of modeling should be biological or physical rules of DNA replication (as in Retkute and Nieduszynski, 2012 [97]) while the NHPP stochastic dynamics should be the output of the modeling. Nevertheless, we feel justified in using a simplified model defined in the terms of the NHPP. One reason is that NHPP can be defined conditionally on the number of replicons and duration of the S-phase, so mathematically our methodology is correct. Another reason is that we use the replication model as a part of a model of cycling cells, to obtain outcome (simulated EdU/DAPI plots) consistent with experiment. This does not contradict the necessity of building de novo models of DNA synthesis, such as those in publications (Hyrien and Goldar, 2010; Guilbaud et al., 2011; Ma et al., 2012; deMoura et al., 2010) [59, 53, 79, 30].

To elucidate whether the observed asymmetry of the EdU vs. DAPI bivariate distributions may be related to limited accessibility of the EdU to DNA replication machinery (and thus to its incorporation) due to a delay in EdU penetration through plasma membrane and/or equilibration with the pool of endogenous dT the point, we carried out an additional experiment: EdU has been added to cells for a short period of 10 sec, and subsequently, 30 min were allowed for the precursor to be incorporated, whereupon the cells were fixed. In this short incubation period, a weak, but characteristic pattern of replication has been already evident (Figure 5.10). It can be thus concluded that EdU penetrates into live cells quickly and efficiently, and is incorporated readily. Admittedly one can argue that it was not possible to wash out all EdU, thus the precursor was slowly entering cells during the 30 minutes period after the main pulse. However, since the cells were thoroughly
Results of the experiment in which EdU has been added to cells for 10 sec and then 30 min were allowed for the precursor to be incorporated, whereupon the cells were fixed. In this short incubation period, a weak, but characteristic pattern of replication has been already present.

rinsed twice we believe that EdU entered cells only during the 10 sec exposure. Furthermore, if the rate of EdU penetration and its equilibration with the endogenous dT plays a significant role in generating the right vs. left asymmetry in the scatterplots, the asymmetry would be much more pronounced during 15 min- and essentially absent during the 120 min- pulse. Clearly, this is not the case. Cells that were entering S during 120 min pulse had ample time to have EdU fully in equilibrium with dT.

It is concluded that even cells that experienced only a part of the EdU pulse (i.e. they started the replication at the end of the pulse or finished during the pulse) accumulated enough EdU for the replication to be detected. They indeed, as we suspected, constitute the relatively large cohort of cells at the left and right bottom ends of the horseshoe scatterplot (Fig. 5.1). Thus the shape of the top (or call it the center) of the horseshoe is not significantly affected by the time required for EdU to enter the cells.
We observed noticeably similar patterns based on comparing DNA labeling distributions between experimental data and simulation results over the course of different exposure times (Fig. 5.1 vs. Fig. 5.8). This is despite of the fact that the replication rate function derived for the simulation shown in Figure 8 was adapted from the replication profile of human embryonic stem cells, whereas the LSC experiments used a human lung cancer cell line (A549). One might expect to observe different DNA replication progression or replication timing control program between normal and cancer cells. However, since they are virtually indistinguishable, this indicates that the time pattern of activation of replication origins is conserved in these different cell lineages. This finding using mathematical modeling with NHPP confirms the observation that the temporal organization of DNA replication sites is almost identical in primary and immortalized mammalian cells (Dimitrova and Berezney, 2002) [35].

In order to understand the consequences of stochastic modeling of DNA replication, we performed a number of simulations to replicate the EdU/DAPI time-course labeling experiments using different replication rate functions. Employing a NHPP in simulating DNA replication progression allows to model the effects of different rates of DNA replication at the early versus late segments of the S-phase, as revealed by the bivariate DNA content distributions. Through a wide range of simulations we demonstrated that the rate function has vital influence on the distribution of labeling intensities. Our simulation software may be used by both specialists and non-specialists to pursue a quantitative understanding of the relation between different cell replication parameters and the observed cell labeling intensities.
In summary, our findings prove consistency between the S-phase DNA-replication rate based on molecular-scale analyses, and cell population kinetics ascertained from EdU/DAPI scatterplots. Our approach opens a possibility of similar modeling to study the effect of anticancer drugs on DNA replication/cell cycle progression and also to quantify other kinetic events that can be measured during S-phase.
Chapter 6

Overall summary and conclusion

The stochastic model and simulation contains random components regarding the behavior of biological phenomena. Research carried out in this dissertation is rooted in the fundamental hypothesis that the dynamics of complex biological systems can be modeled by stochastic processes and stochastic modeling and simulation are more suitable and convenient than deterministic modeling in situations, such as modeling of purely statistical phenomena or if the underlying system is assumed to evolve in pure or semi-random manner or there is lack of available data to form a numerically deterministic model. We investigate four biological applications using stochastic modeling and simulation under two themes. The first theme concerns application of stochastic modeling in genetics, specifically to identify biases in analysis of genetic data. The second theme concerns application of stochastic process models to understand various aspects of cell population dynamics using S-phase labeling. Under both themes, research reveals novel insights into respective biological areas.

In Chapter 2, we investigate the impact of ascertainment bias on interspecies comparisons of microsatellite loci. In genetic inter-species studies, we frequently com-
pare variability at certain locus. Overall, people select high variability locus as the marker locus, the ascertainment bias is rooted at this phenomenon of people tending to choose the locus in such a biased way. After the locus is chosen in species 1, it will be typed in species 2. If we compare the average allele sizes of the two homologous loci in two species, the bias will be reflected as the average allele size in species 2 being smaller than that in species 1. This phenomenon happens in the comparison between human and chimpanzee by choosing a cognate locus in human. The same observation occurs while we compare homologous SNP loci in different human populations using the same biased manner. Furthermore, the reciprocal study shows that ascertainment bias is not always completely due to the sampling effect of choosing marker loci in a biased manner. To study such phenomenon, we propose that the mechanism of ascertainment is a combined effect of sampling and intrinsic difference between the compared species or populations, where sampling effect is due to the tighter correlation of allele sizes within the cognate species than that between two allele with one from species 1 and the other from species 2 and intrinsic difference is due to different demography and genetic factors. Using evolutionary theory we propose a stochastic model to study genetic and sampling effects that influence the comparison of allele size difference, and their impact on estimates of mutation rates. Analytical calculation of the expected allele size difference between two species are validated by forward-time simulation using simuPOP. We re-analyze literature data concerning allele size difference between Human and Chimpanzee. Our results indicate that, when adjusted for ascertainment bias and demographic effects, estimate on microsatellite mutation rate in Human exceeds
or equals that in Chimpanzee. We also demonstrate that changes of population sizes in the recent human history have little impact on conclusions. Our theory and model can be used to explain the apparently discordant conclusions reached by other investigators examining this issue, where the allele size difference could be predominantly due to ascertainment bias alone or the ascertainment bias effect is counteracted or even reversed due to differences in mutation rates and demographic history. It will also aid further research to explore any arbitrarily complex evolutionary scenarios with ascertainment bias and provide a modeling and simulation tool to benefit geneticists in order to estimate genetic and evolutionary parameters in an unbiased manner.

In Chapter 3, we introduce a simulation framework, SimRare, to generate sequence-based data for rare variant association studies under a variety of demographic and disease models, and to evaluate association test methods in an unbiased way. Genetic association studies are performed to determine whether a genetic variant is associated with a phenotype, often a disease or trait. If association is present, a particular allele, genotype or haplotype of a polymorphism or polymorphisms will be seen more often than expected by chance in an individual carrying the trait. Thus, a person carrying one or two copies of a high-risk variant or high-risk genes is at increased risk of developing the associated disease or having the associated trait. Many other factors, such environmental, population structure, sample selection, etc. are confounding genetic associations. In the literature, comparison of rare variant association methods is usually subject to a sampling bias because in any given study data is generated using a test set that makes a particular method
appear superior to others. SimRare is a simulation platform, which generates geno-
type and phenotype data on rare genetic variants (underlying many human diseases) using realistic models, programmed with the aid of simuPOP. SimRare provides an unbiased and easy-to-use platform to evaluate association methods, including novel methods. SimRare is designed to have a user-friendly interface that allows for easy entry of genetic and phenotypic parameters. Novel rare variant association methods that are implemented in R statistical programming language can be imported into SimRare, to compare their performance with existing methods. Using SimRare to compare existing association methods, we demonstrate that the power of each method frequently depends on the test data sets and differences in power among methods are modest. In future research, SimRare can be used as a benchmark program to fairly and easily evaluate rare variant association methods, including novel methods. It can also aid parameter analysis of underlying disease model to test a wide variety of scenarios to achieve adequate power. Furthermore, SimRare will benefit statistical geneticists to facilitate development of novel association methods without implementing special program to generate data in order to evaluate their methods.

In Chapter 4, we build a stochastic model and simulation program using branching processes to represent the adult hippocampal neurogenesis in mice in its early stages, particularly focus on the impact of apoptosis (programmed cell death). Neurogenesis or the formation of new neurons traditionally was believed to occur only during embryonic development. Adult hippocampal neurogenesis, a process of formation of new neurons, is responsible for learning and memory throughout lifetime.
The majority of previous studies have focused on the characteristics of newborn neurons in the late stage. Little is known about the early stages of proliferation and differentiation of neural stem cells and amplification of neural progenitors. Besides, none of previous models considered apoptosis as a crucial factor. Based on the branching process theory, we develop a computational model that represents the early-stage hippocampal neurogenic cascade and allows us to predict the overall efficiency of hippocampal neurogenesis in both normal and diseased conditions. Our primary experimental tool to study neurogenesis is BrdU, an analog of the nucleotide thymidine, which incorporates into the DNA of dividing cells during the S-phase of the cell cycle. Once the cell’s DNA is labeled with BrdU, it will remain labeled and can be detected with specific antibodies. Data were acquired by BrdU pulse-and-chase labeling experiments to trace number of labeled cells of each type throughout a time course in the animal brains, carried in Dr. Maletic-Savatic’s laboratory. Using the model we derive the equilibrium distribution of cell population, simulate the labeling experiments and obtain labeling curves in silico. Fitting simulation results to data allows us to estimate unknown biological parameters, among which the most crucial ones are hitherto unknown rates of apoptosis of different cell stages. Other model parameters are comparable with previously estimated values that were obtained by Encinas et al. (2011). We also show that reducing apoptosis significantly increases adult hippocampal neurogenesis. Our established model and simulation approach to understand early-stage hippocampal neurogenesis will help to predict overall efficiency of hippocampal neurogenesis in both normal and diseased conditions. Parameter estimation can be tuned according to more available
data in the future, which will ultimately provide us with the best possible accuracy to approach experimental conditions aimed to increase the production and survival of newborn neurons.

In Chapter 5, we demonstrate that the experimental asymmetry in DNA-replication scatterplots is consistent with increasing replication initiation rate along the S-phase of cell cycle. In eukaryotic cells, DNA replication begins at multiple points of the genome, the so-called replication origins. Across entire human genome, there are tens of thousands of such origins. However, they are fired at different times and at present little is known about the mechanisms coordinating the timing program. Therefore, understanding kinetics of DNA replication gives an insight into mechanisms revealing specifics of normal and diseased cells proliferation. Stochastic modeling allows relating molecular events to single-cell characteristics assessed by multiparameter cytometry. In the present study we labeled newly synthesized DNA in A549 human lung carcinoma cells with 15 – 120 min pulses of EdU. All DNA was stained with DAPI and cellular fluorescence was measured by laser scanning cytometry. Time-course labeling experiment was performed in Dr. Darzynkiewicz’s laboratory to measure labeling intensities of both DNA content and replication progression over various exposure times. The “horseshoe”-shaped bivariate scatterplots of EdU incorporation vs. DAPI level report the rate of DNA synthesis and the DNA content, respectively. We build a multiscale stochastic model to simulate the synchronized cell population, the DNA replication procedure and cell labeling experiments to reproduce the experimental procedure in silico. For each S-phase cell the time points at which replication origins are fired are modeled by a
non-homogeneous Poisson Process (NHPP). Assuming NHPP rate estimated from independent experiments, simulated EdU/DAPI graphs are nearly indistinguishable from those experimentally observed. This finding proves consistency between the S-phase DNA-replication rate based on molecular-scale analyses, and cell population kinetics ascertained from EdU/DAPI scatterplots. Our approach opens a possibility of similar modeling to study the effect of anticancer drugs on DNA replication/cell cycle progression and also to quantify other kinetic events that can be measured during S-phase.

This dissertation study results not only in discovery of unknown but meaningful biological parameters, but also in four mathematical models and simulation programs that will benefit the research community. Overall results of our research justify the hypothesis stated earlier on (opening paragraph of Chapter 6).
Bibliography


173


