INFORMATION TO USERS

This manuscript has been reproduced from the microfilm master. UMI films the text directly from the original or copy submitted. Thus, some thesis and dissertation copies are in typewriter face, while others may be from any type of computer printer.

The quality of this reproduction is dependent upon the quality of the copy submitted. Broken or indistinct print, colored or poor quality illustrations and photographs, print bleedthrough, substandard margins, and improper alignment can adversely affect reproduction.

In the unlikely event that the author did not send UMI a complete manuscript and there are missing pages, these will be noted. Also, if unauthorized copyright material had to be removed, a note will indicate the deletion.

Oversize materials (e.g., maps, drawings, charts) are reproduced by sectioning the original, beginning at the upper left-hand corner and continuing from left to right in equal sections with small overlaps. Each original is also photographed in one exposure and is included in reduced form at the back of the book.

Photographs included in the original manuscript have been reproduced xerographically in this copy. Higher quality 6" x 9" black and white photographic prints are available for any photographs or illustrations appearing in this copy for an additional charge. Contact UMI directly to order.
RICE UNIVERSITY

Evolution of Microsatellite Loci: Models and Data

by

Leslea Janice Davison

A THESIS SUBMITTED
IN PARTIAL FULFILLMENT OF THE
REQUIREMENTS FOR THE DEGREE

Master of Arts

APPROVED, THESIS COMMITTEE:

Marek Kimmel, Chair
Professor of Statistics
Rice University

Keith A. Baggerly
Assistant Professor of Statistics
Rice University

David Queller
Professor of Ecology and Evolutionary Biology
Rice University

Ranajit Chakraborty
Allan C. King Professor
University of Texas School of Public Health

Houston, Texas
April, 1997
Abstract

Evolution of Microsatellite Loci: Models and Data

by

Leslea Janice Davison

Microsatellite loci are important for genetic studies due to the many desirable qualities they possess. However, speculation continues about the types of mutation that act on microsatellites. In particular, whether different motif types for stable, non-coding loci have different rates of mutation is not clear.

This thesis addresses three aspects of the mutation problem. First, an investigation of the types of mutation acting on microsatellites is presented. The models compared are the single-step stepwise mutation model and the infinite alleles model. Second, two models of allele size distributions under a non-symmetric, single-step SMM model are proposed. Third, an indirect estimate of mutation rates for di- and tetranucleotide loci is obtained.
Acknowledgments

Thanks to the members of my committee, Dr. Marek Kimmel, Dr. Keith Baggerly, Dr. David Queller, and Dr. Ranajit Chakraborty for reading and commenting on my thesis work.

Special thanks to both Dr. Marek Kimmel and Dr. Keith Baggerly, whose guidance and encouragement made completion of this work possible.

Thanks to my fellow graduate students at Rice. Their friendship and humor has been a blessing to me.

Thanks to Rich Lehoucq for listening to, supporting, and encouraging me through the last year. His love and patience have kept me sane through the process of working on this thesis.

And much love and gratitude is due my family, Margaret, John, and Paula Davison, as well as my grandmother, Pearl Dishman. Their gifts of love, encouragement, and sacrifice have made it possible for me to achieve things I never could have achieved on my own.

This research was supported by grants NIH-GM58545 and NSF-DMS9409909.
Contents

Abstract ii
Acknowledgments iii
List of Illustrations vi
List of Tables xii

1 Introduction 1

2 Genetic Background 5
  2.1 Terminology .................................. 5
  2.2 Microsatellite Loci .......................... 6

3 Microsatellite Data from the Genome Database 9
  3.1 The Genome Database ....................... 9
  3.2 Database Errors ............................ 13

4 Models of Mutation 16
  4.1 The Fisher-Wright Model of Random Genetic Drift ........... 17
  4.2 Coalescence .................................. 19
  4.3 Mutation and Microsatellites .................. 21
  4.4 Mutational Mechanisms ....................... 23
4.4.1 Replication Slippage ........................................ 23
4.4.2 Recombinatrial Misalignment During Meiosis ........... 23
4.5 Models of Mutation .............................................. 25
  4.5.1 Infinite Alleles Model .................................. 26
  4.5.2 General Stepwise Mutation Model ................... 28
4.6 Data and Models ............................................... 30
  4.6.1 Estimates of $\theta$ Based on the Infinite Alleles Model and
        Stepwise Mutation Model .................................. 30
4.7 Distributions of Allele Sizes ................................. 56
  4.7.1 Convolution of Poissons Model ....................... 57
  4.7.2 Coalescent Model .................................... 60
  4.7.3 Exploratory Data Analysis ............................. 63

5 Estimation of Relative Mutation Rates of Microsatellite

Loci ......................................................... 85
  5.1 Principles of Estimation ................................. 85
    5.1.1 Relative Mutation Rates Based on GDB Data ....... 89

6 Conclusions ................................................ 93

Bibliography ............................................... 97
Illustrations

4.1 Diagram of replication slippage for a segment of DNA representing a dinucleotide locus. ........................................ 24

4.2 Diagrams of crossing over and recombination during meiosis: (a) homologous chromosomes; (b) replication of homologous chromosomes; (c) crossover event; (d) replicated chromosomes after the completion of the crossover event; (e) resulting gametes. . 25

4.3 Diagram of recombinatorial misalignment for a segment of DNA representing a trinucleotide locus; point of recombination (a); resulting segments after recombination (b). .................................................. 26

4.4 Diagrams of generalized mutation models. ..................... 27

4.5 Quantile-quantile plot of the log variance estimates for GDB dinucleotide data. .................................................. 39

4.6 Scatterplot of biased $\theta$ estimates for dinucleotides under the IAM, least squares line with intercept forced through zero; bootstrapped 90% error bands for the least squares line, and the line of slope one, intercept zero. .............................. 41
4.7 Scatterplot of bias reduced $\theta$ estimates for dinucleotides under the IAM, least squares line with intercept forced through zero; bootstrapped 90% error bands for the least squares line, and the line of slope one, intercept zero.  .............................................. 42

4.8 Scatterplot of biased $\theta$ estimates for dinucleotides under the SMM; scatterplot of log $\theta$ estimates, and the least squares estimate of intercept, $\ln \beta$, when the slope is forced to be one. The second plot also includes 90% error bands for the intercept of the least squares line, and the line of slope one, intercept zero.  .............................................. 43

4.9 Scatterplot of biased $\theta_{SMM,V}$ estimates and bias reduced $\theta_{SMM,H}$ estimates for dinucleotides under the SMM; scatterplot of log $\theta$ estimates, and the least squares estimate of intercept, $\ln \beta$, when the slope is forced to be one. The second plot also includes 90% error bands for the intercept of the least squares line, and the line of slope one, intercept zero.  .............................................. 44

4.10 Empirical cdf plots for biased and bias reduced estimates of $\theta_{IAM,H}$, $\theta_{IAM,n_a}$, and $\theta_{SMM,H}$, dinucleotide loci.  .............................................. 45

4.11 Scatterplot of biased $\theta$ estimates for tetranucleotides under the IAM, least squares line with intercept forced through zero; bootstrapped 90% error bands on the least squares line, and the line of slope one, intercept zero.  .............................................. 49
4.12 Scatterplot of bias reduced $\theta$ estimates for tetranucleotides under the IAM, least squares line with intercept forced through zero; bootstrapped 90% error bands on the least squares line, and the line of slope one, intercept zero. ........................................... 50

4.13 Scatterplot of biased $\theta$ estimates for tetranucleotides under the SMM; scatterplot of log $\theta$ estimates, and least squares estimate of intercept, $\ln \beta$, when the slope is forced to be one. The second plot also includes 90% error bands for the intercept of the least squares line, and the line of slope one, intercept zero. ................................. 51

4.14 Scatterplot of biased $\theta_{SMM,V}$ estimates and bias reduced $\theta_{SMM,H}$ estimates for tetranucleotides under the SMM; scatterplot of log $\theta$ estimates, and least squares estimate of intercept, $\ln \beta$, when the slope is forced to be one. The second plot also includes 90% error bands for the intercept of the least squares line, and the line of slope one, intercept zero. ........................................... 52

4.15 Empirical cdf plots for biased and bias reduced estimates of $\theta_{IAM,H}$, $\theta_{IAM,H}^{ro}$, and $\theta_{SMM,H}$, tetranucleotide loci. ................................. 53

4.16 Example of Convolution of Poisson Distributions, $b = 0.6, d = 0.4$, and $\lambda = 1.5$. .................................................. 60
4.17 Representation of the coalescent tree with mutations added, as described by Theorem 6.10.2 of Roe's thesis. Beginning with two identical alleles (b), descended from a common ancestor in the previous generation (a), we generate a sample by mutating or duplicating an allele after some time \( t_\tau \). When \( n + 1 \) individuals are present in the current sample, throw out the last individual generated, and keep the remaining \( n \) as the sample. ....... 62

4.18 Histograms for D11S908:
\[ \hat{b} = 0.5191, \hat{d} = 0.4809, \hat{\lambda} = 1.3931, n_a = 6, n = 28 \] ....... 68

4.19 Histograms for coalescent simulations related to D11S908:
\[ \hat{b} = 0.5191, \hat{d} = 0.4809, \hat{\theta}_{SM,H} = 8.4720, n = 28 \] ....... 69

4.20 Histograms for coalescent simulations related to D11S908:
\[ \hat{b} = 0.5191, \hat{d} = 0.4809, \hat{\theta}_{SM,V} = 2.8380, n = 28 \] ....... 70

4.21 Histograms for coalescent simulations related to D11S908:
\[ \hat{b} = 0.5191, \hat{d} = 0.4809, \hat{\theta}_{IAM,H} = 2.5242, n = 28 \] ....... 71

4.22 Histograms for coalescent simulations related to D11S908:
\[ \hat{b} = 0.5191, \hat{d} = 0.4809, \hat{\theta}_{IAM,n_a} = 1.5169, n = 28 \] ....... 72

4.23 Histograms for D1S2639: \( \hat{b} = 0.5, \hat{d} = 0.5, \hat{\lambda} = 0.54, n_a = 5, n = 27 \) ....... 74

4.24 Histograms for D2S169:
\[ \hat{b} = 0.6229, \hat{d} = 0.3771, \hat{\lambda} = 0.4844, n_a = 5, n = 56 \] ....... 75
4.25 Histograms for D2S2290:
\[ \hat{b} = 0.0584, \hat{d} = 0.9416, \hat{\lambda} = 0.5304, n_a = 4, n = 28 \] .......................... 76

4.26 Histograms for D3S1317:
\[ \hat{b} = 0.3540, \hat{d} = 0.6460, \hat{\lambda} = 0.9275, n_a = 7, n = 80 \] .......................... 77

4.27 Histograms for D3S3635:
\[ \hat{b} = 0.9849, \hat{d} = 0.01514, \hat{\lambda} = 1.2744, n_a = 6, n = 56 \] .......................... 78

4.28 Histograms for D5S1364:
\[ \hat{b} = 0.8176, \hat{d} = 0.1824, \hat{\lambda} = 0.4337, n_a = 4, n = 74 \] .......................... 79

4.29 Histograms for D7S2434:
\[ \hat{b} = 0.2204, \hat{d} = 0.7796, \hat{\lambda} = 0.4664, n_a = 3, n = 28 \] .......................... 80

4.30 Histograms for D8S531:
\[ \hat{b} = 0.1426, \hat{d} = 0.8574, \hat{\lambda} = 1.4196, n_a = 7, n = 56 \] .......................... 81

4.31 Histograms for D9S117:
\[ \hat{b} = 0.3209, \hat{d} = 0.6791, \hat{\lambda} = 2.1064, n_a = 8, n = 80 \] .......................... 82

4.32 Histograms for D13S1269:
\[ \hat{b} = 0.6519, \hat{d} = 0.3481, \hat{\lambda} = 0.9979, n_a = 6, n = 54 \] .......................... 83

4.33 Histograms for D14S1063:
\[ \hat{b} = 0.7087, \hat{d} = 0.2913, \hat{\lambda} = 0.4139, n_a = 4, n = 56 \] .......................... 84

5.1 Quantile-quantile plot of the log variance estimates for GDB dinucleotide data. ........................................... 88
5.2 Quantile-quantile plot of the log variance estimates for GDB tetranucleotide data. ........................................ 88

5.3 Comparison plot of empirical cumulative distributions for di- and tetranucleotide GDB data on the first fourteen chromosomes. ...... 92
Tables

3.1 Number of loci from each chromosome, by repeat type, that
contain sufficient data to be included in the set of usable data. . . . 11

4.1 Summary of unimodal and nearly unimodal sample distributions
for the fourteen chromosomes sampled. “Almost unimodal” means
that except for a small bump in the center or in the tails of the
sample distribution, it appears unimodal. . . . . . . . . . . . . . . . 65

4.2 Table of test statistics and critical values for a
Kolmogorov-Smirnov test conducted for each sample distribution
associated with a particular locus. The Kolmogorov-Smirnov test
is a test for differences in distributions between a sample and the
theoretical distribution. In this case the theoretical distribution is
the convolution of Poissons. . . . . . . . . . . . . . . . . . . . . . . . 66

4.3 Comparison of average number of alleles generated in simulation
with the observed number of observed from the GDB samples.
The italicized number in each row is the estimate that most closely
approximates the observed number of alleles for the given locus. . . 73
Chapter 1

Introduction

Microsatellites — tandem repeat sequences of two, three, four, or five base pairs — are important in genetic studies due to the many desirable qualities they possess. These loci are highly polymorphic, widely dispersed throughout the genome, and easily identified through polymerase chain reaction (PCR) [26]. Even though microsatellites possess the above qualities, there is continuing speculation about the process(es) by which new allelic types evolve at microsatellite loci. In particular, whether different repeat motif types for stable, non-coding loci have different rates of mutation is not clear. In this thesis we seek to determine if there are different rates for different motif types.

The data used in these analyses are relative allele frequencies, allele sizes in kilobase pairs, and numbers of genes sampled. Because of the sparsity of data for loci with repeat sequences of length three and five (called tri- and pentanucleotides, respectively) these types are not included in the thesis analyses. The data on dinucleotides and tetranucleotides (loci with repeat sequence size of two and four nucleotide base pairs, respectively) are taken from data on the first fourteen chromosomes obtained from the Genome Database (GDB), which is available on the Internet. Data from the remaining nine chromosomes was not collected.
due to time constraints in the collection process. Though data for the last nine chromosomes was not included in this research, the data collected from the first fourteen chromosomes covers approximately 80% of the microsatellite loci. The chromosomes are arranged in size such that the largest chromosome is the first one, and the chromosomes decrease in size to the smallest, chromosome X or Y, depending on the sex of the individual. The data collected from GDB for this research is quite possibly the largest cleaned microsatellite allele frequency data set in existence.

The first analysis presented is an investigation of the compatibility of the multi-step stepwise mutation model, as approximated by the infinite alleles model (IAM), and the single-step stepwise mutation model (SMM) with the data. This is done through indirect methods. One way to compare the two standard population genetics models (IAM and SMM) is through graphical comparison of estimates of a specific parameter, \( \theta \). The value \( \theta \) represents twice the number of mutations per generation present in a population. Both models have more than one estimator that can be derived for \( \theta \), and we make within-model comparisons of these estimators, hoping that estimates of \( \theta \) will be consistent within a model.

Also under investigation in this thesis are specific single-step SMM models of mutation for microsatellite loci. Models currently used to interpret the mutation processes of microsatellite loci may not be the most appropriate models for such mutations, due to overly simplistic model assumptions. For example, the sym-
metric single-step stepwise mutation model introduced by Ohta and Kimura [19] is often used to describe microsatellite mutations. However, research has shown that there may be a bias towards an increase in repeat counts when microsatellites mutate [21, 24, 26]. While we would like to utilize the simplest model that gives a good description of mutation processes, a model that is too simple may not accurately represent these processes. If the current models are not suitable for microsatellites, then new models should be explored.

We propose a new generalized single-step stepwise mutation model to be analyzed, as a less naive alternative to the symmetric single-step stepwise mutation model [19]. By “generalized” we mean a model that allows for either symmetric or asymmetric distribution of mutation probabilities. The parameters of this model are estimated using the method of moments. Because this model assumes single-step mutations, the resulting distribution of allele frequencies will be unimodal. Comparisons are made between observed unimodal frequency distributions from the GDB data and the theoretical distribution under the generalized single-step model.

Under coalescence theory, and assuming a generalized single-step stepwise mutation model, a sample of individuals (genes) can be generated given a) the value of \( \theta \), and b) the probability of losing one repeat if a mutation occurs. Using the estimates of \( \theta \) found above, and the method of moments estimates of the probability under generalized single-step model of decreasing the size of an allele through
mutation by one repeat, we generate simulated distributions of allele sizes. The average number of alleles observed over the 1000 simulations is compared with the observed alleles at a locus in the data. The simulated distributions tend to be irregular in shape, making comparison of distributions difficult.

Finally, we attempt to infer whether different repeat motifs mutate at different rates. The relationship of the genetic variance and the mutation rate is utilized to indirectly estimate the mutation rates of dinucleotides and tetranucleotides. We carry out a Kolmogorov-Smirnov test, to detect if the maximal distance between the empirical cumulative density function for the log of the variance of allele sizes — in terms of number of repeat sequences — is significantly different from what we would expect if the mutation rates of both dinucleotides and tetranucleotides were identical. This test yields a p-value of < 0.001, suggesting that dinucleotides and tetranucleotides do not mutate at the same rate.

Chapter 2 contains a summary of genetic terms necessary to understanding the importance of microsatellite loci. Chapter 3 contains details about the data used in these analyses. This includes information about the Genome Database (GDB) and information about errors in and error correction for the data. In Chapter 4 we examine various models of mutation, and attempt to make inferences about reasonable models for the available data. Chapter 5 is dedicated to estimating relative mutation rates for dinucleotides and tetranucleotides. Finally, Chapter 6 draws conclusions from the results of the above studies.
Chapter 2

Genetic Background

2.1 Terminology

The most basic DNA elements of interest in this study are nucleotides. A nucleotide is a building block of nucleic acid. Four different nucleotides exist: adenine (A), cytosine (C), guanine (G), and thymine (T). We refer to one individual nucleotide as a base, but since DNA is double stranded, nucleotides line up in base pairs (bp). Stable chemical bonds form between adenine and thymine, as well as between cytosine and guanine, making (A-T) and (G-C) — so called Watson-Crick pairs — the typical base pairs.

A specific location of a sequence of nucleotides on a chromosome is called a locus (plural — loci). At a locus, there may exist one or many possible nucleotide sequences. If a locus has multiple possible nucleotide sequences — differing in length of the sequence, or in the nucleotide makeup of the sequence — it is said to be polymorphic. These different realizations of nucleotide sequences at a locus are called alleles. Heterozygosity is a measure of the proportion of individuals (chromosomes) in a population having different alleles at a given locus.

In this thesis, we are interested in microsatellite loci, the alleles of which consist of sequences of 2–5 nucleotide base pairs repeated in tandem. For example,
ACCACCACC... is the beginning of a tandemly repeated sequence with repeat length three. Microsatellites with 2, 3, 4, and 5 base pair repeat units are called *dinucleotides, trinucleotides, tetranucleotides*, and *pentanucleotides*, respectively.

### 2.2 Microsatellite Loci

Microsatellite loci have value in several areas of genetic research. These loci tend to be highly polymorphic, widely dispersed throughout the genome, and easily identified through polymerase chain reaction (PCR) [26]. The microsatellites that have been studied most thoroughly thus far — dinucleotides — occur at least once every 100,000 base pairs [25]. Because of these properties, microsatellites are becoming important for forensic tests, genome mapping, and population studies.

Forensic studies are conducted to identify individuals through comparison of DNA samples. Because microsatellites are highly polymorphic, they are useful for determining whether two DNA samples came from the same individual. The greater the heterozygosity a microsatellite locus exhibits, the greater the probability of differentiating between individuals if in fact DNA samples do not match.

Microsatellite loci, because of their propensity to have several alleles, may be included in a genome mapping study. In genome mapping, one attempts to estimate the relative location of a gene on a given chromosome. If two genes are relatively close together on a chromosome, they tend to travel together when recombination
occurs. A locus with a greater number of alleles, such as a microsatellite locus, makes it easier to detect recombination events between loci.

For population studies, comparing allele size distributions at microsatellite loci for different populations or even different species may yield information about the evolutionary history of these populations [13]. For example, different species may have different ranges of allele sizes (in terms of number of repeat sequences).

Certain trinucleotide loci have been associated with the hereditary diseases such as myotonic dystrophy, fragile X syndrome, and Huntington’s disease. For these trinucleotide loci, a certain range of repeats is considered normal, and the genes are stable. However, when these genes have higher than normal repeat counts, they are much more mutable as they are passed from one generation to the next [21].

Ashley and Sherman [3] provide information on the normal, premutation, and full mutation allele sizes for fragile X syndrome. The number of repeats for a normal allele for the fragile X site falls in the range of 6–50 [3]. A “normal” allele sequence for fragile X is made up of several CGG repeats, with 0–3 single AGG sequences interspersed approximately every ten CGG repeats. Individuals carrying the premutation form of the gene have alleles consisting of 50–200 CGG repeats. These individuals exhibit no symptoms of fragile X syndrome, but this gene is unstable and more prone towards expansion in future generations. The premutation and full mutation genes, as compared to their normal counterpart,
do not contain the necessary additional AGG sequences sprinkled through the sequence to stabilize the gene. This deficiency appears to cause the instability that makes these genes expand rapidly. Individuals with a full mutation form of the gene have repeat counts exceeding 200, and exhibit mild to severe mental retardation, as well as physical abnormalities.

As microsatellites become more important in many areas of genetic research, understanding their mutational mechanisms and the relationships between mutation rates of different motif types is imperative.
Chapter 3

Microsatellite Data from the Genome Database

3.1 The Genome Database

The Genome Database (GDB) is an Internet site that houses and curates genome mapping data [23]. These data are provided by various researchers who are involved in the Human Genome Project, a research effort committed to mapping the estimated 100,000 human genes. At the time the data for this research were extracted from GDB, version 5.6 was the working version. In January 1996, version 6.0 of GDB was introduced. Features of this new version include allowing public depositing and editing of new data, allowing the user to annotate previously stored data, and allowing the user to customize map queries. No updates to the data contained in GDB 5.6 have been made since the advent of GDB 6.0. Because the data were in a compact format and easy to obtain using the GDB Browser available with version 5.6, we collected all data from version 5.6. Our compilation of microsatellite data on the first fourteen chromosomes is likely the largest cleaned data set of its kind in existence.

Many types of database searches are possible using the GDB Browser (through version 5.6). The browser lists the following as topical searches: polymorphism,
map, probe, library, contact (researcher reporting the data), cell line, citation, loci, mutation, and OMIM (the Online Mendelian Inheritance in Man database).

Using a GDB polymorphism query, we collected di-, tri-, tetra-, and pentanucleotide data from presumably neutral loci in a Caucasian population on the fourteen largest chromosomes. On these chromosomes, there were 4507 out of 5171 dinucleotide loci with usable frequency information, 6 out of 395 for trinucleotides, 46 out of 1995 for tetranucleotides, and only one non-coding pentanucleotide locus, D14S566, had complete, accurate data of the type required for this study. Table 3.1 summarizes the complete data for our data set. Section 3.2 gives an explanation of what errors eliminate a locus from use in our study. The data for dinucleotides were both more extensive and more complete than for any of the other three motif types. The dinucleotide sequence \((C - A)_n (G - T)_n\) is the most frequently identified simple sequence in the human genome [27].

Within a polymorphism query, many types of information are provided. For this work, data for non-gene associated microsatellite loci was collected. Of the data gathered, the allele sizes in kilobase pairs, the relative frequencies of different allele sizes, the number of chromosomes sampled, and the number of distinct alleles observed are of greatest interest. Also recorded (when available) were the researcher(s) reporting the data, the publication where the data appeared, the map location of a locus on a chromosome, and the heterozygosity for each locus.
<table>
<thead>
<tr>
<th>Chromosome</th>
<th>Number of Loci with Useful Information</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Di-</td>
</tr>
<tr>
<td>1</td>
<td>477</td>
</tr>
<tr>
<td>2</td>
<td>468</td>
</tr>
<tr>
<td>3</td>
<td>413</td>
</tr>
<tr>
<td>4</td>
<td>311</td>
</tr>
<tr>
<td>5</td>
<td>375</td>
</tr>
<tr>
<td>6</td>
<td>338</td>
</tr>
<tr>
<td>7</td>
<td>298</td>
</tr>
<tr>
<td>8</td>
<td>290</td>
</tr>
<tr>
<td>9</td>
<td>249</td>
</tr>
<tr>
<td>10</td>
<td>317</td>
</tr>
<tr>
<td>11</td>
<td>299</td>
</tr>
<tr>
<td>12</td>
<td>280</td>
</tr>
<tr>
<td>13</td>
<td>213</td>
</tr>
<tr>
<td>14</td>
<td>179</td>
</tr>
<tr>
<td><strong>Totals</strong></td>
<td><strong>4507</strong></td>
</tr>
</tbody>
</table>

Table 3.1 Number of loci from each chromosome, by repeat type, that contain sufficient data to be included in the set of usable data.
Investigation of the relationship between individuals from whom chromosome samples were extracted shows that more than 80% of the dinucleotide loci in GDB have data from samples of unrelated individuals. Three researchers provided more than 80% of the dinucleotide data: Jean Weissenbach (> 65%), James Weber (≈ 5%), and Mark Gyapay (≈ 12%). All three authors cite CEPH (le Centre d’Etudes du Polymorphisme Humain) as the source of their data. CEPH is responsible for maintaining data from 61 large reference families, for the purpose of creating genetic maps [1]. The samples provided by CEPH for use by GDB are from unrelated Caucasian individuals among the 61 pedigrees.

Among the other ≈ 20% of the dinucleotide data, each individual contributor provided less than 1% (less than 45 loci). In many cases, a contributor may only be the author cited for one locus in all 14 chromosomes analyzed. A survey conducted via email and by checking papers referenced in the GDB database of a few other data providers suggests that most of the loci in GDB are from samples of unrelated individuals. Dr. Michael Dixon’s samples (for 10 loci in GDB) were from unrelated individuals. Dr. David Kwiatkowski’s samples (for 23 loci in GDB) were from Venezuelan reference pedigrees. The number of related individuals in these samples is less than 10%. Samples used by Dr. V. Orphanos were from 30 unrelated individuals, including 18 individuals from 8 CEPH families.

For the forty-six tetranucleotide loci with complete data that were retained for analysis, it was verified that the chromosome samples were from unrelated individ-
uals for thirty-nine of the loci. This verification was done by referring to articles cited for each locus in GDB. For seven of the loci, reported by K.P. Petrukhin, there was no literature reference cited. Attempts at personal communication via email with Dr. Petrukhin yielded no results.

3.2 Database Errors

Inconsistencies concerning the quantity and accuracy of data reported to GDB exist due to the varying interests of researchers reporting the data. For example, the data for locus D1S1181 in GDB were reported by S. C. Gerken [23]. While Gerken includes information about the number of chromosomes sampled (30) and the probe used to locate this dinucleotide sequence, the author does not include any other information that would be informative for our study. Incidentally, Gerken’s entries in general included only probe and chromosome sample size information. This lack of data is attributable not only to Gerken, but to other researchers as well. The sample for locus D6S260 reported by J. Weissenbach has eighteen alleles listed, but relative frequency values for seven of the alleles are missing. For locus D3S1672, M. Wei reports observation of nine alleles, with frequency values for the nine recorded; yet the sum of the frequencies of the nine alleles is only 0.29 — that is only 29% of the sample is accounted for. Whether there is a data entry error in the frequency values or there are alleles not reported is unclear. Problems such as those suggested in the examples above are not uncommon in the GDB data.
Another database error is round-off error of relative frequencies. The relative frequency values corresponding to different allele sizes have been rounded in different ways by different data contributors. In many cases, relative allele frequencies are simply rounded to two decimal places. This rounding may cause the overall sum of allele frequencies for a given locus to yield sums of just over or under one. In some cases, relative allele frequencies have been truncated to two decimal places. This results in sums of relative frequencies of less than one. This error is corrected with the following process: 1) multiply the relative frequency of each allele by the number of chromosomes in the sample; 2) take the ceiling of this value; 3) divide this by the number of chromosomes in the sample; 4) round to two decimal places. The resulting value is taken as the relative frequency value for the given allele.

In some cases, the sums of frequency values are not close to one, making it necessary to decide upon cut-off values such that a sum of frequencies over or under these cut-off values would result in discarding the locus from the data set. The values 0.90 and 1.10 were arbitrarily set as the inclusive lower and upper bounds for retaining data. For cases where rounding to two decimal places introduced error, frequency estimates have been standardized by dividing each allele frequency by the sum of all of the allele frequencies for that locus.

In some cases the number recorded as the number of chromosomes in the sample appears to be equal to the number of individuals contributing chromosomes in the sample, based on the fact that some of the reported frequencies correspond to
fractional numbers of chromosomes (e.g. $< 1/2$ chromosome). When such a case is detected, the number of chromosomes is simply multiplied by two, taking into account that an individual has two copies of each chromosome.

In cases where the number of chromosomes sampled for a locus was not reported, the locus was retained. For most cases, the number of chromosomes sampled is a tool used for unbiasing estimates for different estimators. If this value is unavailable, we have simply settled for biased estimates.

Loci that had missing frequency or allele size values have been eliminated from the analysis.

GDB contains the largest set of microsatellite loci data available. However, errors within the data make it necessary to screen the data before using it in analysis. With the data we obtained for the first fourteen chromosomes in the human genome, we are able to draw many conclusions about the nature of mutations for microsatellites.
Chapter 4

Models of Mutation

In this chapter we cover several aspects of the mutation process and its affect on microsatellites. First we introduce the Fisher-Wright model of mutation and the related coalescent model. Next we discuss two mutational mechanisms believed to be acting on microsatellites: replication slippage and recombinatorial misalignment during meiosis. Various models of mutation are then introduced. The two models we are interested in exploring are the infinite alleles model (IAM) and the single-step stepwise mutation model (SMM). The IAM is used as an estimate of the multi-step SMM when mutations are restricted to additions or deletions of repeats.

Using the estimates of a parameter $\theta$ under both the IAM and single-step SMM, we try to make inferences about the type of mutations acting on microsatellites. Through estimates of $\theta$ we may infer that either single-step mutations dominate in the mutation process of microsatellites, or that mutations involving the loss or gain of more that one repeat are quite prevalent. These inferences are made for dinucleotides and tetranucleotides based on data collected from the Genome Database (GDB).

Finally, we explore two distributions of allele sizes that depend on different assumptions about dependencies between individuals within a population. One
model assumes that the size of the Caucasian population is large enough that dependencies in allele sizes among individuals are negligible. This model is based on the propagation of mutations over time under a Markov Chain model. The other model is based on the assumption that some dependency exists between individuals in the Caucasian population. Simulated samples of allele sizes corresponding to samples for certain loci in GDB are generated using a coalescent model. This coalescent simulation depends on the estimate of the parameter $\theta$ and the estimated single-step mutation probabilities for the corresponding sample from GDB.

4.1 The Fisher-Wright Model of Random Genetic Drift

To understand mutation, we must have a model which illustrates the forces causing it to occur. Two such models, the Fisher-Wright and coalescent models, are presented below.

The Fisher-Wright model (also known as the Wright-Fisher model) of random genetic drift is the basis of many population genetics studies today. This model uses a Markov-Chain formulation to describe the allele frequency distribution for populations subject to random genetic drift. For this model, assume we begin with a subpopulation of effective size $N$ from an infinite population. If we apply this theory to a human population, we see that for $N$ diploid individuals, we have $2N$ haploid copies of each chromosome. In the first generation there exist $k$ alleles at a particular locus, each allele having relative frequency $p_i \ (i \in 1\ldots k)$ in the
subpopulation. However, due to random sampling, this relative frequency will not remain fixed through the generations. To create the next generation (which is non-overlapping with the first generation), we sample with replacement $2N$ times from the current generation, resulting in a constant population size in all generations. Just as flipping a fair coin 100 times most likely will not yield exactly 50 heads and 50 tails, neither will the resampling scheme for each generation yield exactly the same relative frequencies of each allele as were observed in the previous generation. In fact, the smaller the subpopulation (that is, the smaller $N$ is), the more rapidly the locus will move towards being fixed for one allele, in general. This means that after this resampling scheme is followed for many generations, one allele will become more and more dominant in the subpopulation while other alleles decrease in frequency. Then at some time $t$ in the future, that one allele will be the only one in existence for its locus. This move towards fixation is what is commonly referred to as genetic drift.

Now assume that for a human population (with $N$ individuals and $2N$ copies of each chromosome) there are two alleles, $A_1$ and $A_2$, at a given locus. Then we can describe the state of the Fisher-Wright model as the number of alleles of type $A_1$ (which in turn specifies the number of alleles of type $A_2$) existing in the population, where the state space is the non-negative integers up to $2N$. In this two allele situation we may describe the transition probabilities with the binomial distribution. Thus the transition from $i$ copies of allele $A_1$ in one generation to $j$
copies of this allele in the following generation has probability

\[ p_{ij} = \binom{2N}{j} \left( \frac{i}{2N} \right)^j \left( 1 - \frac{i}{2N} \right)^{2N-j}. \]

This theory may be generalized for \( m \) alleles at a locus, where genotypes — all possible pairings of these alleles — are described by a multinomial distribution. This model may also be altered to include the effects of mutation. In such a case, states 0 and 2\( N \), where the population is fixed for allele \( A_1 \) or \( A_2 \), respectively, are not absorbing states, since mutation may bring allelic variability back into the population. If the combination of mutation and drift eventually create a steady state distribution for allele sizes, this state is aptly called mutation-drift equilibrium.

For more in depth treatment see Hartl and Clark [11] or Ewens [10].

4.2 Coalescence

Another important theory associated with the Fisher-Wright model is that of coalescence. The theory of coalescence is that for any sample of size \( n \) from the 2\( N \) chromosomes in a population, those \( n \) individuals have a single common ancestor at some time in the past. To trace the coalescence of any two alleles in this sample of size \( n \), we see that there are \( \binom{n}{2} \) ways to choose the two alleles that will coalesce first from the current sample. Then in the generation of complete coalescence of the \( n \) currently existing individuals of interest (\( t \) generations ago), each of the 2\( N \) chromosomes has a \( 1/2N \) probability of being the ancestor from
which the two alleles chosen in the current generation descended. Thus for some continuous time $T_n$, the time when there are $n$ individuals existing in the population with no common ancestor, $T_n$ has an exponential distribution with parameter $\binom{n}{2}/2N$. Continuing back in time, $T_{n-1}$, the time before $n-1$ individuals coalesce in some previous generation to $n-2$ individuals, has an exponential distribution with parameter $\binom{n-1}{2}/2N$. Eventually, the 2 remaining individuals coalesce to 1, and the exponential distribution of this coalescent time has parameter $1/2N$.

Assuming small population size, the relevance of coalescent theory for the study of mutations in microsatellite loci is due to dependencies between individuals in a relatively small population. By using the coalescent tree as a framework, mutation models which incorporate such dependency as exists in small populations can be formulated. We now apply coalescence theory to a sample of size $n$ at a microsatellite locus collected from $N$ individuals. Recall that $N$ individuals contribute $2N$ alleles at a locus to a population. Let the random variable $X$ have the distribution of allele sizes for a particular microsatellite locus. By coalescence theory, we know that since the $n$ individuals in the sample are descendants of one common ancestor in the past, their allele sizes at a microsatellite locus cannot be independent. However, the sizes are exchangeable; the joint distribution of the $X_i, i = 1, 2, \ldots, n$ is the same for all permutations of the $X_i$.

When effective population size is large enough or population expansion is rapid, we may treat the dependencies amongst individuals as trivial, and apply simpler
theoretical models. In this case, we develop a model in the form of expected frequencies. Given the assumption of independence, we may compare the sample frequencies to the expected frequency distribution. This is because the sample frequencies should well approximate the expected frequencies as the sample size gets large if the model is valid. Further description of the coalescent simulation as well as the derivation of expected frequency distribution are contained in Section 4.7.

4.3 Mutation and Microsatellites

One of the ways loci avoid fixation is mutation. In the context of this thesis, mutation is defined as a change in allelic state at a given locus, by the addition or deletion of one or more repeats to the sequence, and mutation rate refers to the expected number of mutations at a locus in a given generation. A subpopulation is said to be mutation-drift balanced if random genetic drift eliminates alleles at a particular locus from a subpopulation at the same rate as mutation introduces new or previously observed alleles into the subpopulation. This balance prevents fixation for a particular allele at a locus, and maintains a steady-state distribution of allele sizes for the given locus.

According to direct studies, mutation rates observed for microsatellite loci are relatively high. Mutation rates in general are on the order of $10^{-6}$ to $10^{-4}$ mutations per locus per generation [11], but studies have shown that microsatellite
loci have mutation rates on the order of $10^{-3}$ mutations per locus per generation [18, 26, 27].

Weber and Wong (1993) [26] define mutation rate as

$$\nu \equiv \frac{\text{Number of mutation events}}{\text{Number of Meioses}}.$$

Under ideal circumstances, mutation rates are estimated by direct observation of mutation events occurring in study families. By "ideal circumstances" we mean being able to observe actual mutation events in an extremely large sample of individuals. The difficulties with direct estimation are that mutations are rare, and sample size – that is, the number of participants in family studies – is typically small. Thus observation of a single mutation may have a large effect on the estimate of mutation rate. If mutational "hot spots" exist in the sample of loci, these may bias an estimate of the mutation rate. Instead of relying on direct observations of mutations, we approach the problem of estimating the mutation rate indirectly. By making use of data on the distribution of allele sizes, we will estimate the mutation rates for different repeat motifs of microsatellite loci. This is described in detail in Section 5.
4.4 Mutational Mechanisms

The physical mutation mechanisms of microsatellite loci are not yet well understood. The two mutational mechanisms believed to be in operation are replication slippage and recombinatorial misalignment [21].

4.4.1 Replication Slippage

Replication slippage occurs when through some misalignment in the replication process, a "bubble" is created and one or more repeat sequences do not have complementary nucleotides synthesized for them (see Figure 4.1). In such a case this bubble is excised, resulting in a plasmid — a piece of DNA that is not attached to any particular chromosome. This can result in the loss of a repeat. Alternatively, this plasmid might attach itself to the complementary strand of DNA for the chromosome being replicated, resulting in the addition of a repeat [20]. Current research suggests this may be the most common type of mutation acting on microsatellites.

4.4.2 Recombinational Misalignment During Meiosis

The other potential source of mutations which we consider is the recombinatorial misalignment that may occur during meiosis, in cells that form gametes. This process is also known as unequal sister chromatid exchange. Each cell in a human body contains two copies of each chromosome — one passed down from that
human's mother and one from the father. These two copies (called a diploid set) are homologous: both copies of a chromosome contain the same genes, but not necessarily identical alleles for each gene.

During the early part of the process of forming gametes, each chromosome in the cell replicates, forming two identical strands of DNA joined together at an area called the centromere. When joined in this way, the replicated chromosomes have a cross-like structure. The two identical joined strands are termed sister chromatids. Chromatid arms from homologous pairs may cross over one another, and exchange sequences (Figure 4.2, adapted from Biology Hypertext [8]). If unequal amounts of information are exchanged between swapped arms, new alleles not present in either of the parents will be the result (Figure 4.3).
This mutational mechanism can be differentiated from replication slippage because flanking markers are swapped in recombinatorial misalignment, but not in replication slippage.

4.5 Models of Mutation

In population genetics, two models used to simulate mutation are the the infinite alleles model (IAM), and the stepwise mutation model (SMM) (see Figure 4.4).

Figure 4.2 Diagams of crossing over and recombination during meiosis: (a) homologous chromosomes; (b) replication of homologous chromosomes; (c) crossover event; (d) replicated chromosomes after the completion of the crossover event; (e) resulting gametes.
Figure 4.3 Diagram of recombinatorial misalignment for a segment of DNA representing a trinucleotide locus; point of recombination (a); resulting segments after recombination (b).

4.5.1 Infinite Alleles Model

Under the infinite alleles model, each chromosome is independently subject, with probability \( \nu \), to a mutation where an allele of size \( X \) is replaced by a new allele not previously seen in the population. In this case, a mutation may not be restricted to addition or subtraction of whole repeats, but instead include any type of mutation possible; two examples of other types of mutation would be base substitutions and sequence inversions. Given the types of mutation that are thought to occur for microsatellites, this model of mutation does not appear to be useful for modelling microsatellite mutation. However, if attention is restricted to mutations of the addition or subtraction of repeats type, then a IAM is the limiting distribution of
the SMM. That is, the closer to uniform on the integers the distribution of $U$ is, the more the SMM resembles the IAM.

Figure 4.4 Diagrams of generalized mutation models.
4.5.2 General Stepwise Mutation Model

Under the stepwise mutation model, each chromosome is independently subject, with probability \( \nu \), to a mutation which replaces an allele of size \( X \) (number of repeats) with an allele of size \( X + U \), where \( U \) is defined to be an integer-valued random variable with a specified distribution. Examples of feasible distributions for \( U \) are given below. This model seems appropriate for describing microsatellite mutation, since a step may be defined as the loss or gain of a certain number of repeats. Several special cases of the stepwise mutation model are discussed below.

Symmetric Single-Step Stepwise Mutation Model

The simplest SMM, Ohta and Kimura’s symmetric single-step SMM, assumes that a mutation results in the gain or loss of exactly one repeat, and that there is equal probability of gain or loss [19]. If the mutation rate is \( \nu \), then the probability of gaining (losing) a repeat is \( \nu/2 \), and the probability of no mutation is \( 1 - \nu \). The single-step model has been used by many researchers to make inferences about microsatellite loci [22, 24, 26]. Levinson and Gutman [17] conducted experiments that showed replication slippage mutation to be of one repeat unit in most cases.

Symmetric Two-Step and Two-Sided Geometric Mutation Models

Kimmel and Chakraborty [13] put forward the symmetric two-step and two-sided geometric as two other examples of feasible models of microsatellite mutation. The
symmetric two-step mutation model has probabilities $\nu p/2$ for gain or loss of one repeat, and $\nu(1 - p)/2$ for gain or loss of two repeats. In a generalized version of this model, the two steps could be of absolute size one or $k$, with $k$ having possible values $2, 3, 4, \ldots$. It is assumed that if a mutation occurs under the two-step model, $p \gg (1 - p)$; that is, a mutation of one step is much more likely than a mutation of more than one step.

Giving $U$ a two-sided geometric distribution allows mutations to be of all integer valued step sizes. However, mutational probabilities in this model must be scaled to take into account that probability $(1 - q)/(1 + q)$ is assigned to mutations of size zero. The sum over all non-zero mutations is then $2q/(1 + q)$ instead of one, so we must multiply all probabilities of non-zero mutation by $(1 + q)/2q$ to normalize the probability measure. For the parameter $q \in (0, 1)$ of the two-sided geometric, $q$ near zero gives results very similar to those for the one- and two-step stepwise mutation models, and $q$ near one yields a uniform distribution on the integers (excluding zero) [13].

Generalized Single-Step Stepwise Mutation Model

In this thesis, a generalized single-step stepwise mutation model is explored. Research has shown that there may be a bias towards an increase in repeat counts when a microsatellite mutates [21, 24, 26]. With this model, loss and gain of one repeat have probabilities $\nu d$ and $\nu b$, respectively, where $b + d = 1$. Although the occur-
rence of mutations of absolute size change greater than one repeat is feasible, we will assume these mutations to be rare enough to be negligible. This model will be discussed in further detail in Section 4.7.1.

4.6 Data and Models

4.6.1 Estimates of $\theta$ Based on the Infinite Alleles Model and Stepwise Mutation Model

In order to assess which model best describes the observed data, we would like a method of relating the data to the two standard models of mutation, the IAM and SMM introduced in Section 4.5. We use the IAM as an approximation to the multi-step SMM, because the IAM is more tractable. One way we can make comparisons is through estimating some parameter common to both models. For this work, we will estimate the composite parameter, $\theta = 4N\nu$, which represents twice the equilibrium number of mutations per generation at a given locus. The parameter $\nu$ as discussed in Section 4.3 is the mutation rate for a particular locus. The parameter $N$ represents the effective population size of diploid individuals. The parameter $\theta$ may also be defined as $\theta = 2N\nu$ if we instead let $N$ represent the number of haploid individuals (that is, chromosomes) in the effective population. This representation will be used in Section 4.7.2 for the coalescent model. The effective population size tends to be smaller than the actual population size, taking into consideration such things as unequal number of males and females in the
population, and age structuring of the population which affects the number of individuals participating in reproduction [11].

**Theoretical Expressions**

Under the IAM and the single-step SMM, introduced in Section 4.5, we explore two equations for each model that contain the parameter $\theta$. Under the IAM, the equilibrium heterozygosity at a locus is expressed as

$$H = \frac{\theta}{\theta + 1}$$

(see Hartl and Clark, Chapter 3 [11]). Also according to this model, the expected number of alleles at equilibrium is expressed as

$$E(n_a) = \sum_{i=0}^{n-1} \frac{\theta}{\theta + i},$$

where $n$ is the number of chromosomes sampled (see Ewens [9]).

Under the single-step SMM, the equilibrium heterozygosity equation is

$$H = 1 - \frac{1}{\sqrt{1 + 2\theta}}$$

(see Ohta and Kimura [19]), and the equation for genetic variance at equilibrium is

$$\frac{V}{2} = \frac{\theta}{2}$$

(see Kimmel et al. [15] and Kimmel and Chakraborty [13]).
To calculate estimates of \( \theta \) we must first obtain estimates of heterozygosity, number of alleles, and genetic variance at a locus. Heterozygosity is defined as the proportion of diploid individuals in a population having different alleles at a given locus. Thus the equation to represent heterozygosity is

\[
H = 1 - \sum_{i=1}^{n_a} p_{ii}.
\]

Here \( n_a \) is the number of alleles present at the given locus, and \( p_{ii} \) is the relative frequency of homozygotes \( ii \) in the population. To estimate heterozygosity, we use the observed number of alleles, \( \hat{n}_a \), as our estimate of the true number of alleles at a locus, and the relative frequencies of the alleles in the sample, \( k_i/n \), as the estimate of the \( p_i \), the true frequency of the \( i^{th} \) allele in a population of haploid individuals. If we can assume independence of individuals, then \( p_{ii} = (k_i/n)^2 = p_i^2 \).

Thus the estimate of heterozygosity is

\[
\hat{H} = \frac{n}{n-1} \left[ 1 - \sum_{i=1}^{n_a} \left( \frac{k_i}{n} \right)^2 \right],
\]

where \( k_i \) represents the number of chromosomes in the sample having allele type \( i \). It can be shown that \( E(\hat{H}) = 1 - \sum_{i=1}^{n_a} p_{ii} \), so this estimator of heterozygosity is unbiased.

For the estimator of \( \theta \) based on the expression for the expected number of alleles, we replace the expected number with the observed number of alleles, \( \hat{n}_a \), at the locus of interest.
For the estimator of $\theta$ based on the genetic variance, we estimate the variance with the usual variance estimator,

$$\hat{V} = \frac{1}{n-1} \sum_{j=1}^{n} (X_j - \overline{X})^2,$$

where $n$ is the number of chromosomes sampled, and $X_j$ is the allele size (in number of repeats) of the $j^{th}$ chromosome in the sample.

Because we are interested in estimates of $\theta$, we solve the above IAM and single-step SMM equations for $\theta$, and plug in the necessary estimate of heterozygosity, number of alleles, or variance as previously described. The four estimates of $\theta$ are denoted as follows: $\hat{\theta}_{IAM,H}$ is the IAM estimator based on heterozygosity; $\hat{\theta}_{IAM,na}$ is the IAM estimator based on the expected number of alleles equation; $\hat{\theta}_{SMM,H}$ is the single-step SMM estimator based on heterozygosity; and $\hat{\theta}_{SMM,V}$ is the single-step SMM estimator based on variance.

The $\hat{\theta}_{IAM,na}$ estimator is a non-negative, real solution of the expected number of alleles expression

$$\hat{n}_a = \sum_{i=0}^{n-1} \frac{\hat{\theta}_{IAM,na}}{\hat{\theta}_{IAM,na} + i}.$$

Taking into consideration the process of sampling alleles at a locus, we need to allow for the fact that different researchers have different priorities when recording the frequencies of alleles sampled. For instance, researchers doing genotyping may discard alleles that appear only once in the sample, assuming they are errors on a gel. Stuttering of lines that appear on a gel may occur, making it unclear if a
rare allele, observed on a gel, is real or is a mistake. Thus some of the loci in GDB may be biased towards only alleles that appear more than once, when the loci are provided by researchers whose interest is in genotyping. To correct for this, we may use an estimator for the number of alleles that appear only once in a sample [4].

Chakraborty [4] proposed an formula to estimate the number of rare alleles in a sample at a locus. The formula for the expected number of alleles that appear in only one copy is

$$E(n_{(a,1)}) = \frac{n\theta}{n + \theta - 1}.$$  

A bias reduced estimate of $\theta$ based on the number of alleles observed in the sample may be reached by combining the above two equations, and solving numerically for $\hat{\theta}_{IAM,n_a}$:

$$0 = \sum_{i=0}^{n-1} \frac{\hat{\theta}_{IAM,n_a} + i}{n + \hat{\theta}_{IAM,n_a} - 1} - \left( \hat{n}_a - \hat{n}_{(a,1)} \right),$$

where $n$ is the sample size, $\hat{n}_a$ is the number of alleles observed in the sample, and $\hat{n}_{(a,1)}$ is the number of alleles that appear in only one copy in the sample.

The $\theta_{IAM,H}$ estimator is

$$\hat{\theta}_{IAM,H} = \frac{\hat{H}}{1 - \hat{H}}.$$  

Note that this estimate is biased, and unbiasing is a difficult task in this case.

Chakraborty and Weiss [7] proposed a bias corrected estimate of $\theta_{IAM,H}$. The
corrected value is obtained by numerically solving

\[ \hat{\theta}^2_{IAM,H} + (7 - \theta^*)\hat{\theta}^2_{IAM,H} + (8 - 5\theta^*)\hat{\theta}_{IAM,H} - 6\theta^* = 0 \]

for \( \hat{\theta}_{IAM,H} \), using the estimate obtained before the bias correction, \( \theta^* \), as a starting value. An equivalent direct solution may be obtained by solving for \( \hat{\theta}_{IAM,H} \) in the Chakraborty and Weiss equation and inputing the estimate \( \theta^* \). This requires a mathematical package such as Maple.

For the single-step SMM the resulting \( \theta \) estimators are

\[ \hat{\theta}_{SMM,H} = \frac{1}{2} \left[ \frac{1}{(1 - \hat{H})^2} - 1 \right] \]

and

\[ \hat{\theta}_{SMM,V} = \hat{V}. \]

Again, both estimators are biased. Kimmel, Chakraborty, and Jorde ([14], paper in progress) propose a bias reduced estimator of \( \theta_{SMM,H} \) based on theory developed by Zouros [29]. Let \( \hat{H} \) represents the sample heterozygosity for a locus, and

\[ E(\hat{H}) = 1 - \frac{1}{\sqrt{1 + 2\theta}}, \]

\[ Var(\hat{H}) \approx \frac{\theta}{3 + 11.25\theta + 13\theta^2 + 1.7\theta^3}. \]

Then a bias corrected estimate of \( \theta \) may be obtained through application of the Taylor series approximation relationship

\[ E[f(\hat{H})] \approx f\left[E(\hat{H})\right] + f''\left[E(\hat{H})\right] Var(\hat{H})/2, \]
where \( f(x) = [(1 - x)^{-2} - 1]/2 \). So we see that \( f(\hat{\theta}) = \theta^* \) is the biased estimate of \( \theta \) obtained above. Also note that \( f'[E(\hat{\theta})] = \theta \), and \( f''[E(\hat{\theta})] = 3(1 - E(\hat{\theta}))^{-4} = 3(1 + 2\theta)^2 \). We plug these values into the Taylor series approximation to get

\[
E(\theta^*) \approx \theta + \frac{3\theta(1 + 2\theta)^2}{6 + 22.5\theta + 26\theta^2 + 3.4\theta^3}.
\]

If we substitute the biased estimate \( \theta^* \) obtained above for \( E(\theta^*) \), and substitute \( \hat{\theta} \) for \( \theta \), we arrive at the equation

\[
3.4\hat{\theta}^4 + (38 - 3.4\theta^*)\hat{\theta}^3 + (34.5 - 26\theta^*)\hat{\theta}^2 + (9 - 22.5\theta^*)\hat{\theta} - 6\theta^* = 0.
\]

Using Maple, the above equation may be explicitly solved to get the estimate \( \hat{\theta} \) from a function of \( \theta^* \). If \( \theta^* \) is positive, then only one of the four roots of this equation is both positive and real valued. This is the estimate that we retain as our bias corrected estimate of \( \theta_{SMM,H} \).

The estimates of \( \theta_{SMM,V} \) are biased, and no equation for bias reduction of the \( \theta_{SMM,V} \) equation has been proposed.

Using the data we have obtained for each microsatellite locus, we arrive at estimates of the above quantities — heterozygosity, expected number of alleles, and genetic variance — and then use these quantities for estimating \( \theta \) values to get comparable estimates of the number of mutations occurring at a locus. We calculate and show results for both biased and unbiased estimates. Unfortunately the data is not without flaw (see Chapter 3), so some adjustments must be made to be able to estimate \( \theta \) values. Certain loci do not have the number of chromosomes
sampled recorded. Consequently, \( \hat{\theta}_{IAM,n_a} \) could not be estimated for these loci. The \( \hat{\theta}_{SMM,V} \) estimator also depends on the number of chromosomes sampled, to make the estimate unbiased. We chose to include slightly biased estimates in these cases rather than remove these loci from our analysis. For estimating heterozygosity and variance, relative frequencies of allele sizes for each loci are required. In cases where the sum of the relative frequencies is less than one due to round-off error, we standardize the frequencies by dividing the frequency of each allele by the sum of the frequencies.

Exploratory Data Analysis

We now compare estimates of \( \theta \) based on allele count/variability to those based on heterozygosity. If for a given repeat motif the population is at equilibrium, then the number of alleles present at a locus should be "enough" for the corresponding heterozygosity. Estimates of \( \theta \) were obtained for both the dinucleotides and the tetranucleotides on the first fourteen chromosomes. Trinucleotides did not contain enough complete data to be included in this study (see Chapter 3). Scatterplots of the \( \hat{\theta}_{IAM,n_a} \) values against the \( \hat{\theta}_{IAM,H} \) values and \( \hat{\theta}_{SMM,V} \) against \( \hat{\theta}_{SMM,H} \) were produced for the dinucleotides and the same procedure was followed for the tetranucleotides. This comparison is made for both biased and unbiased estimates.
The scatterplots of the $\hat{\theta}_{SM, V}$ versus $\hat{\theta}_{SM, H}$ for both dinucleotides and tetranucleotides show the estimates to be very dispersed. Chakraborty et al. [5] concluded that the large variability of $\hat{V}$ suggests a highly skewed sampling distribution, reflecting large variation within repeat motif for the given set of loci. A quantile-quantile plot of the natural log variance estimates for the dinucleotide data show that the natural log variances are very close to normally distributed (see Figure 4.5). Therefore, a natural log transformation was applied to the $\theta_{SM, V}$ and $\theta_{SM, H}$ estimates providing a better method of examining the fit of the single-step SMM to the data. The log of $\theta_{SM, H}$ is taken so that we may compare the two SMM estimators on the same scale. Assuming that the relationship of the $\theta$ estimators is explained by a linear model

$$\theta_{SM, V} = \beta \cdot \theta_{SM, H} + \varepsilon,$$

where $\beta$ is the slope, $\varepsilon$ is the error term, and the intercept is forced through zero, taking natural logs gives us a new linear model,

$$\ln \theta_{SM, V} = \ln \beta + \ln \theta_{SM, H} + \eta.$$

Minimizing the following sum of squared deviations equation with respect to $\ln \beta$,

$$\sum_{i=1}^{t} [\ln \theta_{SM, V, i} - (\ln \beta + \ln \theta_{SM, H, i})]^2$$

where $t$ is the number of loci observed, we get the estimate

$$\ln \hat{\beta} = \ln \hat{\theta}_{SM, V} - \ln \hat{\theta}_{SM, H}.$$
Exponentiating \( \ln \hat{\beta} \) returns a less biased estimate of the slope of the linear relationship of \( \theta_{SMM,V} \) and \( \theta_{SMM,H} \), when the intercept is forced through zero.

If \( \theta \) is estimated consistently within a model, then a plot of \( \theta \) estimates, estimated in two different ways under one model, should have a linear relationship, with slope one and intercept zero. Plots roughly demonstrating these attributes suggest the model is generally appropriate for loci of the given motif type plotted.

To test whether a least squares line of slope one and intercept zero is feasible for the IAM given the data, we bootstrapped 100 samples of size equivalent to the number of theta coordinates included in the plot. For each sample we fitted a least

![Quantile-quantile plot for Log Variances, Dinucleotides](image)

**Figure 4.5** Quantile-quantile plot of the log variance estimates for GDB dinucleotide data.
squares line, and then took the 5th smallest and 5th largest slopes from the 100 generated, to get a 90% error band for the least squares line.

For the log transformed single-step SMM estimates of $\theta$, we similarly constructed 90% error bands for the intercept, $\ln \beta$. We chose the 5th smallest and 5th largest intercept estimates from 100 bootstrapped samples to construct the 90% error bands. By exponentiating the estimated 5th and 95th quantiles, we get the error bands for the slope of the untransformed $\theta$ estimates.

The first plot in Figure 4.6 shows the scatterplot of the biased estimates of $\hat{\theta}_{IAM,n_a}$ versus the biased estimates of $\hat{\theta}_{IAM,H}$ for dinucleotides with the least squares line through the data. The second plot shows the error bands for the least squares line, as well as the line of slope one and intercept zero. The first plot shows that the estimates are well concentrated, and have a fairly linear relationship. A least squares line fitted to the data, intercept forced through zero, has slope 0.85. In the second plot, the bootstrapped 90% error bands on the least squares line are quite narrow, and do not contain the line with slope one and intercept zero. The bootstrapped confidence interval for the slope is (0.8251, 0.8674). The correlation between the biased estimates of $\hat{\theta}_{IAM,H}$ and $\hat{\theta}_{IAM,n_a}$ is 0.73.

Figure 4.7 shows the scatterplot of the bias reduced estimates $\hat{\theta}_{IAM,n_a}$ versus the bias reduced estimates $\hat{\theta}_{IAM,H}$ for dinucleotides, and the error bands for the least squares line through the data. The estimates are well concentrated, and have a fairly linear relationship. A least squares line fitted to the data, intercept forced
Figure 4.6 Scatterplot of biased $\theta$ estimates for dinucleotides under the IAM, least squares line with intercept forced through zero; bootstrapped 90% error bands for the least squares line, and the line of slope one, intercept zero.
Figure 4.7 Scatterplot of bias reduced \( \theta \) estimates for dinucleotides under the IAM, least squares line with intercept forced through zero; bootstrapped 90% error bands for the least squares line, and the line of slope one, intercept zero.
Figure 4.8  Scatterplot of biased $\theta$ estimates for dinucleotides under the SMM; scatterplot of log $\theta$ estimates, and the least squares estimate of intercept, $\ln \beta$, when the slope is forced to be one. The second plot also includes 90% error bands for the intercept of the least squares line, and the line of slope one, intercept zero.
Figure 4.9 Scatterplot of biased $\theta_{SMM,V}$ estimates and bias reduced $\theta_{SMM,H}$ estimates for dinucleotides under the SMM; scatterplot of log $\theta$ estimates, and the least squares estimate of intercept, $\ln \beta$, when the slope is forced to be one. The second plot also includes 90% error bands for the intercept of the least squares line, and the line of slope one, intercept zero.
Figure 4.10 Empirical cdf plots for biased and bias reduced estimates of \( \theta_{IAM,H} \), \( \theta_{IAM,na} \), and \( \theta_{SMM,H} \), dinucleotide loci.
through zero, has slope 1.15. Bootstrapped 90% error bands for the least squares line are quite narrow, and do not contain the line with slope one and intercept zero. The bootstrapped confidence interval for the slope is (1.1100, 1.1816). The correlation between $\hat{\theta}_{IAM,H}$ and $\hat{\theta}_{IAM,n_a}$ is 0.71.

The top plot in Figure 4.8 shows the scatterplot of the biased estimates $\hat{\theta}_{SM, V}$ versus the biased estimates $\hat{\theta}_{SM, H}$ for dinucleotides. These estimates are very dispersed, showing little linear relationship. The correlation between $\hat{\theta}_{SM, H}$ and $\hat{\theta}_{SM, V}$ is 0.36.

Figure 4.8 also shows the scatterplot of the log-transformed biased estimates of $\theta_{SM, V}$ and $\theta_{SM, H}$ for the dinucleotides. The least squares line for the data, calculated as explained above, has intercept $(\ln \hat{\beta}) = -5.5 \cdot 10^{-5}$, and the slope is forced to be one. The 90% error bands for the intercept are extremely narrow, and contain the line of slope one and intercept zero. The bootstrapped confidence interval for the intercept is (-0.0245, 0.02686). Exponentiating $\ln \beta$ yields $\hat{\beta} = e^{(-5.5 \cdot 10^{-5})} = .999945 \approx 1$. This demonstrates that the relationship between $\theta_{SM, V}$ and $\theta_{SM, H}$, when the bias of the variance estimator is corrected, is linear with slope $\approx 1$ when the intercept is forced through zero. Exponentiating the 5% and 95% estimated cutoff values gives us the estimated confidence interval for the slope of the untransformed estimates (0.9758, 1.0272). The line of slope one and intercept zero falls between these bands. The correlation between the transformed estimates is 0.61.
Figure 4.9 shows the scatterplot of the biased estimates $\hat{\theta}_{SMM,V}$ versus the bias reduced estimates $\hat{\theta}_{SMM,H}$ for dinucleotides. These estimates are very dispersed, with little linear relationship. The correlation between $\hat{\theta}_{SMM,H}$ and $\hat{\theta}_{SMM,V}$ is 0.34.

Figure 4.9 also shows the scatterplot of the log-transformed biased estimates of $\theta_{SMM,V}$ and the bias reduced estimates of $\theta_{SMM,H}$ for the dinucleotides. The least squares line for the transformed data has intercept $(\ln \hat{\beta}) = 0.27$, and the slope is forced to be one. Error bands for the intercept are very narrow, with a bootstrapped confidence interval for the intercept of $(0.2537, 0.2980)$. This interval does not contain the intercept zero. Exponentiating $\ln \hat{\beta}$ we get $\hat{\beta} = e^{0.27} = 1.31$. This demonstrates that the relationship between $\theta_{SMM,V}$ and $\theta_{SMM,H}$, when the bias of the variance estimator is corrected, is linear with slope $\approx 1.31$ when the intercept is forced through zero. The bootstrapped confidence interval for the slope of the untransformed $\theta$ estimates is $(1.2888, 1.3471)$, which does not contain the slope one. The correlation between the transformed estimates is 0.63.

Comparisons of the empirical cumulative density functions for the biased and bias reduced estimates of $\theta$ for $\theta_{IAM,H}$, $\theta_{IAM,n_a}$, and $\theta_{SMM,H}$ for dinucleotides appear in Figure 4.10. The curve for the bias reduced estimates of $\theta_{IAM,H}$ falls to the left of the curve for the biased estimates. This means that the bias reduction process for estimates of $\theta_{IAM,H}$ uniformly produces smaller estimates than the biased counterpart. For the estimates of $\theta_{IAM,n_a}$, the two cdfs are quite close together and overlap, and one does not uniformly fall to the left of the other.
The curve for the bias reduced estimates of $\theta_{SMM,H}$ falls slightly to the left of the curve for the biased estimates, suggesting that bias reduction procedure produces slightly smaller estimates of $\theta_{SMM,H}$.

Figure 4.11 shows the scatterplot of the biased $\hat{\theta}_{IAM,n_a}$ versus biased $\hat{\theta}_{IAM,H}$ for tetranucleotide loci. The estimates have a strong linear relationship, where a least squares line has slope 0.55 when the intercept is forced through zero. The correlation is 0.82. The 90% bootstrapped confidence interval for the slope of the regression line, (0.4986, 0.6129), does not contain the slope one.

Figure 4.12 shows the scatterplot of bias reduced $\hat{\theta}_{IAM,n_a}$ versus bias reduced $\hat{\theta}_{IAM,H}$ for tetranucleotide loci. The estimates have a strong linear relationship, where a least squares line has slope 0.74 when the intercept is forced through zero. The correlation is 0.63. The bootstrapped 90% confidence interval for the slope of the regression line, (0.6019, 0.9086), does not contain the slope one.

Figure 4.13 shows the scatterplot of the biased estimates of $\theta_{SMM,V}$ and the biased estimates of $\theta_{SMM,H}$ for tetranucleotide loci. These estimates, like their dinucleotide counterparts under the SMM, are quite dispersed. The correlation between $\hat{\theta}_{SMM,H}$ and $\hat{\theta}_{SMM,V}$ is 0.60.

Figure 4.13 also shows the scatterplot of the log-transformed biased estimates of $\theta_{SMM,V}$ and the log-transformed biased estimates of $\theta_{SMM,H}$ for the tetranucleotides. The estimates demonstrate a better linear relationship than the untransformed estimates, yet the estimates are still fairly dispersed. Correlation between
Figure 4.11 Scatterplot of biased $\theta$ estimates for tetranucleotides under the IAM, least squares line with intercept forced through zero; bootstrapped 90% error bands on the least squares line, and the line of slope one, intercept zero.
Figure 4.12 Scatterplot of bias reduced $\theta$ estimates for tetranucleotides under the IAM, least squares line with intercept forced through zero; bootstrapped 90% error bands on the least squares line, and the line of slope one, intercept zero.
Figure 4.13  Scatterplot of biased $\theta$ estimates for tetranucleotides under the SMM; scatterplot of log $\theta$ estimates, and least squares estimate of intercept, $\ln \beta$, when the slope is forced to be one. The second plot also includes 90% error bands for the intercept of the least squares line, and the line of slope one, intercept zero.
Figure 4.14 Scatterplot of biased $\theta_{SMM,V}$ estimates and bias reduced $\theta_{SMM,H}$ estimates for tetranucleotides under the SMM; scatterplot of log $\theta$ estimates, and least squares estimate of intercept, $\ln \beta$, when the slope is forced to be one. The second plot also includes 90% error bands for the intercept of the least squares line, and the line of slope one, intercept zero.
Figure 4.15 Empirical cdf plots for biased and bias reduced estimates of \( \theta_{IAM,H} \), \( \theta_{IAM,na} \), and \( \theta_{SMM,H} \), tetranucleotide loci.
\( \theta_{SMM,V} \) and \( \theta_{SMM,H} \) is 0.81. Using the method described above, we obtained the estimate \( \ln \hat{\beta} = -0.84 \). The 90% bootstrapped confidence interval for the intercept is \((-1.0255, -0.6557)\), and it does not contain the slope one. Exponentiating \( \ln \hat{\beta} \) we get \( \hat{\beta} = 0.43 \). This suggests that estimate of \( \theta_{SMM,H} \) is on average 2.32 times greater than the estimate of \( \theta_{SMM,V} \). The bootstrapped confidence interval for the slope, when the transformed estimates are exponentiated, is \((0.3586, 0.5191)\).

Figure 4.14 shows the scatterplot of biased \( \theta_{SMM,V} \) ands bias reduced \( \theta_{SMM,H} \) estimates for tetranucleotide loci. These estimates are quite dispersed. The correlation between \( \hat{\theta}_{SMM,H} \) and \( \hat{\theta}_{SMM,V} \) is 0.59.

Figure 4.14 also shows the scatterplot of the log-transformed biased estimates of \( \theta_{SMM,V} \) and the log-transformed bias reduced estimates of \( \theta_{SMM,H} \) for the tetranucleotides. The estimates demonstrate a better linear relationship than the untransformed estimates, yet the estimates are still fairly dispersed. Correlation between \( \theta_{SMM,V} \) and \( \theta_{SMM,H} \) is 0.81. The estimate \( \ln \hat{\beta} = -0.59 \). The bootstrapped confidence interval for the intercept is \((-0.7465, -0.4078)\). The intercept zero does not fall in this interval. Exponentiating \( \ln \hat{\beta} \) we get \( \hat{\beta} = 0.55 \). This implies that the estimates of \( \theta_{SMM,H} \) are on average 1.82 times greater than the estimates of \( \theta_{SMM,V} \). The bootstrapped confidence interval for the slope, when the transformed error band estimates are exponentiated, is \((0.4740, 0.6651)\).

Comparisons of the empirical cumulative density functions for the biased and bias reduced estimates of \( \theta \) for \( \theta_{IAM,H}, \theta_{IAM,n}, \) and \( \theta_{SMM,H} \) for tetranucleotides...
appear in Figure 4.15. The curves for the bias reduced estimates of $\theta_{IAM,H}$ and $\theta_{SMM,H}$ fall to the left of the curves for the biased estimates. This means that the bias reduction process uniformly produces smaller estimates than the more biased estimation counterpart. For the estimates of $\theta_{IAM,n_a}$, the two cdfs are quite close together and overlap, and one does not uniformly fall to the left of the other.

The above analysis shows that for the dinucleotide data, estimates of $\theta$ under the IAM are fairly consistent. When biased estimators of $\theta_{IAM,H}$ and $\theta_{IAM,n_a}$ are compared, they have a linear relationship with a slope of approximately 0.85. The bias reduced estimators yield estimates that have a linear relationship with slope of approximately 1.15. The linear relationships observed in both the biased and bias reduced estimates are fairly near one, and the biased and bias reduced linear relationships vary about the line of slope one.

Under the single-step SMM applied to the dinucleotide data, the biased estimates of $\theta_{SMM,H}$ and $\theta_{SMM,V}$ were quite consistent. The estimated intercept of the natural log transformed estimates of $\theta$ was approximately $5.5 \cdot 10^{-5}$. This translates into an estimate of the slope of $\approx 1$ for the linear relationship between $\theta_{SMM,H}$ and $\theta_{SMM,V}$. For the bias reduced estimates, the estimated slope of the linear relationship between $\theta_{SMM,H}$ and $\theta_{SMM,V}$ is $\approx 1.31$. The results of the application of the IAM and the single-step SMM estimators of $\theta$ to the dinucleotide data suggest that because both the IAM and the single step SMM yield consistent estimates of $\theta$ within a model. However, the consistency demonstrated by both
models suggest that we have no power to differentiate between the IAM and the single-step SMM for dinucleotide loci using the method of comparing estimates of $\theta$ under the two models.

For the tetranucleotide data, $\hat{\theta}_{IAM,n_a}$ and $\hat{\theta}_{SMM,V}$ are systematically lower than estimates of $\theta$ based on heterozygosity ($\hat{\theta}_{IAM,H}$ and $\hat{\theta}_{SMM,H}$, respectively). The biased estimates of $\theta_{IAM,H}$ and $\theta_{IAM,n_a}$ have a linear relationship, but with a slope of approximately 0.55. The bias reduced estimates of $\theta_{IAM,H}$ and $\theta_{IAM,n_a}$ have a linear relationship, with slope of approximately 0.74. The estimated slopes of the linear relationship between $\theta_{SMM,H}$ and $\theta_{SMM,V}$ estimates, calculated through the process of taking natural logs of the $\theta$ estimates, also are not near one. The estimated slope of the linear relationship between biased estimates of $\theta_{SMM,V}$ and biased estimates of $\theta_{SMM,H}$ was 0.43. The estimated slope of the linear relationship between biased estimates of $\theta_{SMM,V}$ and bias reduced estimates of $\theta_{SMM,H}$ was 0.55. Although all these estimates show a linear relationship, the estimate of slope markedly lower than one suggests a model different from equilibrium IAM or SMM would be more appropriate.

4.7 Distributions of Allele Sizes

In this section we explore two types of distributions that are both relevant — given different population size assumptions — to the explanation of the mutation mechanisms of microsatellite loci. If we assume that the population is small, then
mutational dependencies may exist amongst the individuals in the population. On the other hand, if the population size is quite large, these possible dependencies amongst individuals may be not be so strong, and thus can be ignored in modeling. We base a model for the first case on the theory of the coalescence, outlined in Section 4.2. A model for the large population case can be based on the expected frequencies of allele sizes because of the minimal dependence we assume between individuals. We develop such a model based on a generalized single-step SMM. This thesis does not explore multi-step models due to the added difficulty in estimating parameters. Because of this limitation, we only compare the results of this model with unimodal sample allele size distributions, which are assumed to have mutated under single-step stepwise mutation.

4.7.1 Convolution of Poissons Model

Assuming we have a large, expanding population, a coalescent tree for a sample of that population would have extremely long branches representing the times to coalescence events. That is to say the coalescent events of this sample are in the remote past for the existing population, making the allele sizes of individuals in the sample nearly independent. Because of this near independence, we can create a distribution based on following the mutations of the lineage of one allele created in the distant past to represent the distribution of allele sizes at an equilibrium state, conditional on this lineage not dying out. The expected allele frequencies
evolve in time as a Markov chain, with the mutation process unaffected by genetic drift. Thus for the parameters

\[ \nu = \text{mutation rate}, \]

\[ b = \text{probability of gaining one repeat unit, given mutation occurs}, \]

\[ d = 1 - b = \text{probability of losing one repeat unit, given mutation occurs}, \]

we can create a transition matrix for the allelic state at a locus:

\[
p_{ij} = \begin{cases} 
\nu d, & j = i - 1, \\
1 - \nu, & j = i, \\
\nu b, & j = i + 1, \\
0, & |j - i| > 1.
\end{cases}
\]

If the \( U_i \) are iid random variables representing the mutational distribution for a single time period, \( U_i \in \{-1, 0, 1\} \), a chain starting at state 0 has the characteristic function

\[ \Phi_{U_i}(s) = E(e^{istU_i}) = \nu de^{-js} + (1 - \nu) + \nu be^{js}, \]

after a single generation. We are interested in the distribution of allele sizes after the passing of \( t \) generations, which amounts to the accumulation of independent mutations over \( t \) time intervals

\[ S_t = \sum_{i=1}^{t} U_i. \]

This yields a characteristic function of the form

\[ \Phi_{S_t}(s) = E(e^{isS_t}) \]
\[ = \left[ \nu de^{-js} + (1 - \nu) + \nu be^{js} \right]^t \]
\[ = \left[ 1 - \frac{tvd(1 - e^{-js}) + tvb(1 - e^{js})}{t} \right]^t. \]

In such a case, as

\( t \) gets large,

\( \nu \) approaches zero, and

\( tv \to \lambda, \)

the above characteristic function is approximated by

\[ \exp\{\lambda b(e^{is} - 1)\} \cdot \exp\{\lambda d(e^{-js} - 1)\}. \]

This is a convolution of Poisson distributions, where one Poisson distribution has parameter \( \lambda b \) and the other Poisson distribution is reflected about the origin, with parameter \( \lambda d \):

\[ \text{Poisson}(\lambda b) \ast \text{Poisson}_{(-)}(\lambda d). \]

Probabilities for this type of convolution of Poissons are specified as

\[ \phi_j = e^{-\lambda} \left( \frac{b}{d} \right)^j I_j(2\lambda\sqrt{bd}) \quad \phi_j, \ j \in \mathbb{Z}, \]

where \( I_j \) is the \( j^{th} \) order modified Bessel function of the first kind [2].

Estimates of the parameters \( b \) (and thus \( d \)) and \( \lambda \) are obtained through method of moments estimation. The parameter estimates are

\[ \hat{b} = \frac{(1 + \hat{m}_3/\hat{m}_2)}{2}, \]
\[ \hat{\lambda} = \hat{m}_2, \]
where \( m_k \) is the \( k^{th} \) central moment from the sample.

For example, if \( b = 0.6, \ d = 0.4, \) and \( \lambda = 1.5, \) then the convolution of Poisson distributions would have the distribution shown in Figure 4.16.

### 4.7.2 Coalescent Model

Using the theory of coalescence outlined in Section 4.2, Theorem 6.10.2 of Roe's doctoral dissertation [20] describes a method of generating the joint distribution of a sample of size \( n \) from a population of size \( N, \) as \( N \to \infty, \) where \( N \) represents the effective population size of haploid individuals. Note that this is different from notation in the previous sections. In Roe's thesis, a genealogy is created for a given sample of \( n \) individuals, and mutations are superimposed onto this genealogy. Figure 4.17 gives a pictorial explanation of the process of generating the sample. This method allows the branches of the tree to evolve independently. The type

![Figure 4.16 Example of Convolution of Poisson Distributions, \( b = 0.6, d = 0.4, \) and \( \lambda = 1.5. \)]
of mutation is restricted to non-recombinatorial mutations, since a recombination event requires the mutation of two "individuals" (chromosomes), and this would introduce dependence between individuals. Assumptions of constant sized populations through the generations hold. The parameter $\theta$ is equivalent to the previous definition in this thesis. Below is the theorem as it appears in Roe's thesis:

**Theorem 6.10.2** An urn scheme for modelling sequence evolution. (Donnelly and Kurtz (1993).)

A robust description of the [allele] types in a sample of size $n$ from a population evolving according to a general exchangeable model for evolution with (in rescaled time) Markov mutation process of rate $\theta/2$, ... is given by the following urn scheme.

1. Start at [time] step $\tau = 2$ with two identical [allele] sequences $\varepsilon_1 = \varepsilon_2$, whose [allele] type is a realization from the stationary distribution of the (continuous time) Markov mutation process.

2. If at stage $\tau$ there are $m$ sequences, then at the next stage duplicate a particular sequence

   $$(\varepsilon_1, \ldots, \varepsilon_i, \ldots, \varepsilon_m) \rightarrow (\varepsilon_1, \ldots, \varepsilon_i, \ldots, \varepsilon_m, \varepsilon_i)$$

   with probability $\frac{m-1}{m(m-1+\theta)}$, $i = 1, 2, \ldots, m$

   or mutate an allele sequence

   $$(\varepsilon_1, \ldots, \varepsilon_i, \ldots, \varepsilon_m) \rightarrow (\varepsilon_1, \ldots, \varepsilon_i^*, \ldots, \varepsilon_m)$$

   with probability $\frac{\theta}{m(m-1+\theta)}$, $i = 1, 2, \ldots, m$

   where $\varepsilon_i^*$ is a mutant offspring of $\varepsilon_i$ whose distribution is

   $$\Pr(\varepsilon_i^* = \mu) = p_{\varepsilon_i, \mu}.$$  

3. Continue in this way until the first occasion on which there are $n + 1$ sequences and then discard the last sequence.

The remaining sequences $(\varepsilon_1, \ldots, \varepsilon_n)$ have the distribution of a sample of size $n$ from a population of size $N$, in the limit as $N \to \infty$, evolving according to a general exchangeable model of evolution with mutation.
Figure 4.17  Representation of the coalescent tree with mutations added, as described by Theorem 6.10.2 of Roe’s thesis. Beginning with two identical alleles (b), descended from a common ancestor in the previous generation (a), we generate a sample by mutating or duplicating an allele after some time $t_r$. When $n + 1$ individuals are present in the current sample, throw out the last individual generated, and keep the remaining $n$ as the sample.
The equation $\Pr(\epsilon_i^* = \mu) = p_{\epsilon_i^*,\mu}$ represents the probability that the mutant allele is of type $\epsilon_i^*$ conditional on the parent having allele type $\epsilon_i$.

Based on this theorem, King [16] created Fortran code to generate samples of size two from a population of size $N$. King used this code to estimate the genetic variance at a locus. We modified the code to generate a sample of size $n$ under a non-symmetric single-step SMM, given the parameters $\theta, d$ (the probability of losing one repeat sequence given mutation occurs), and a starting value for the pseudo-random number generator. For this analysis we chose a selection of twelve loci with unimodal or near unimodal distributions randomly chosen from the fourteen chromosomes surveyed. The parameters $\theta$ and $d$ were estimated using ad hoc methods: $\theta$ estimates were those based on estimates of heterozygosity, allele count, and variance under the IAM and single-step SMM, described in Section 4.6.1; $d$ was obtained using the method of moments estimator for the convolution of Poissons model, as described in Section 4.7.1. While the coalescent model assumes non-symmetric, single-step stepwise mutation, we also generated distributions using the estimates of $\theta$ obtained under the IAM for comparison.

### 4.7.3 Exploratory Data Analysis

Histograms were generated to compare the convolution of Poissons model to GDB sample distributions for selected unimodal loci. A non-parametric Kolmogorov-Smirnov test was then conducted for each unimodal sample versus the convolu-
tion of Poisson distributions. Also for each locus, an average of the number of alleles observed in each of the 1000 coalescent simulations was compared with the number of alleles observed in the sample. These two comparisons were done for twelve unimodal dinucleotide loci, randomly selected from the fourteen chromosomes: D1S2639, D2S169, D2S2290, D3S1317, D3S3635, D5S1364, D7S2434, D8S531, D9S117, D11S908, D13S1269, D14S1063. On the 14 chromosomes sampled, there are 433 unimodal sample distributions out of the 4507 loci collected. In addition, there are approximately 150 near unimodal distributions. This means that the distribution may have a slight bump for a certain allele, but otherwise appears unimodal. A detailed breakdown of the unimodal distributions on the fourteen chromosomes sampled appears in Table 4.1.

To create a histogram for the convolution of Poissons to compare with a sample distribution, method of moments estimates of the parameters $b, d$, and $\lambda$ are calculated from the sample distribution of allele sizes. Then probabilities are calculated for each allele size under a convolution of Poisson distributions with the estimated values of $b, d$, and $\lambda$. If in the tails of the resulting distribution the probability for a given allele size is less than 0.01, we combine this probability with the nearest bin having probability $\geq 0.01$. Histograms for the convolution of Poissons model and for the sample distribution are then generated.

A non-parametric Kolmogorov-Smirnov test for differences in distributions was conducted for the estimated convolution of Poisson distributions and the sam-
<table>
<thead>
<tr>
<th>Chromosome</th>
<th>Unimodal</th>
<th>Almost Unimodal</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>49</td>
<td>11</td>
</tr>
<tr>
<td>2</td>
<td>48</td>
<td>18</td>
</tr>
<tr>
<td>3</td>
<td>46</td>
<td>17</td>
</tr>
<tr>
<td>4</td>
<td>40</td>
<td>8</td>
</tr>
<tr>
<td>5</td>
<td>38</td>
<td>8</td>
</tr>
<tr>
<td>6</td>
<td>19</td>
<td>9</td>
</tr>
<tr>
<td>7</td>
<td>17</td>
<td>10</td>
</tr>
<tr>
<td>8</td>
<td>31</td>
<td>11</td>
</tr>
<tr>
<td>9</td>
<td>26</td>
<td>8</td>
</tr>
<tr>
<td>10</td>
<td>25</td>
<td>14</td>
</tr>
<tr>
<td>11</td>
<td>36</td>
<td>15</td>
</tr>
<tr>
<td>12</td>
<td>20</td>
<td>4</td>
</tr>
<tr>
<td>13</td>
<td>23</td>
<td>11</td>
</tr>
<tr>
<td>14</td>
<td>15</td>
<td>9</td>
</tr>
<tr>
<td><strong>Totals</strong></td>
<td><strong>433</strong></td>
<td><strong>153</strong></td>
</tr>
</tbody>
</table>

Table 4.1  Summary of unimodal and nearly unimodal sample distributions for the fourteen chromosomes sampled. “Almost unimodal” means that except for a small bump in the center or in the tails of the sample distribution, it appears unimodal.
<table>
<thead>
<tr>
<th>Locus</th>
<th>Test Statistic</th>
<th>Critical Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>D1S2639</td>
<td>0.0371</td>
<td>0.1705</td>
</tr>
<tr>
<td>D2S169</td>
<td>0.0732</td>
<td>0.1184</td>
</tr>
<tr>
<td>D2S2290</td>
<td>0.5657</td>
<td>0.1674</td>
</tr>
<tr>
<td>D3S1317</td>
<td>0.1816</td>
<td>0.0991</td>
</tr>
<tr>
<td>D3S3635</td>
<td>0.2520</td>
<td>0.1184</td>
</tr>
<tr>
<td>D5S1364</td>
<td>0.0368</td>
<td>0.1030</td>
</tr>
<tr>
<td>D7S2434</td>
<td>0.1700</td>
<td>0.1674</td>
</tr>
<tr>
<td>D8S531</td>
<td>0.0762</td>
<td>0.1184</td>
</tr>
<tr>
<td>D9S117</td>
<td>0.1028</td>
<td>0.0991</td>
</tr>
<tr>
<td>D11S908</td>
<td>0.0681</td>
<td>0.1674</td>
</tr>
<tr>
<td>D13S1269</td>
<td>0.0448</td>
<td>0.1206</td>
</tr>
<tr>
<td>D14S1063</td>
<td>0.2138</td>
<td>0.1184</td>
</tr>
</tbody>
</table>

Table 4.2 Table of test statistics and critical values for a Kolmogorov-Smirnov test conducted for each sample distribution associated with a particular locus. The Kolmogorov-Smirnov test is a test for differences in distributions between a sample and the theoretical distribution. In this case the theoretical distribution is the convolution of Poissons.

ple distribution for each of the twelve unimodal loci. The null hypothesis of the test is that the distribution of the sample is the same as the theoretical distribution. The test statistic is based on the greatest absolute difference in the cumulative distribution functions, and takes into consideration the sample size of the sample distribution. An asymptotic approximation for the test statistic for the Kolmogorov-Smirnov test at the $\alpha = 0.05$ level is $0.886/\sqrt{n}$ where $n$ is the sample size. The results of the Kolmogorov-Smirnov tests on the twelve loci are as follows: for D1S2639, D2S169, D5S1364, D8S531, D11S908, and D13S1269, we declined to reject the null hypothesis that the sample distribution differed from the theor-
ical convolution of Poisson distributions; for D2S2290, D3S1317, D3S3635, and D14S1063, we rejected the null hypothesis; for D7S2434 the test statistic is 0.1700 and the critical value is 0.1674; and for D9S117 the test statistic is 0.1028 and the critical value is 0.0991; thus for D7S2434 and D9S117 we make no inference. The test statistics and critical values for each locus are listed in Table 4.2.

Figure 4.18 shows the histograms for the sample and for the theoretical convolution of Poissons for locus D11S908. Figures 4.19 – 4.22 show histograms for randomly selected coalescent simulations for the given $\hat{\theta}$ values. The coalescent simulation histograms show that distributions with a range of shapes and number of alleles are possible under the simulation conditions. Another important feature is that almost none of the distributions simulated from the coalescent have the regularity of the sample distribution.

Using the 1000 simulated samples generated for each locus, we calculate an average number of alleles present in the 1000 simulations. Because of the irregularity of the histograms for the simulations, this average number of alleles is a means to compare the simulated samples with the sample from GDB. Table 4.3 shows the average under the different estimates of $\theta$, for each of the twelve loci. The italicized number in each row is the estimate that best approximates the number of alleles observed. For all loci except D1S2639 and D7S2434, the simulations created with the $\theta_{SSM,H}$ estimates were the closest approximations to the number of alleles observed in the corresponding sample from GDB. In the case of D1S2639,
Sample plot for Dinucleotide locus D11S908

Theoretical distribution under convolution of poisson's model

Figure 4.18  Histograms for D11S908:
\( \hat{b} = 0.5191, \hat{d} = 0.4809, \lambda = 1.3931, n_a = 6, n = 28 \)
Figure 4.19  Histograms for coalescent simulations related to D11S908: \( \hat{b} = 0.5191, \hat{d} = 0.4809, \hat{\theta}_{SM,H} = 8.4720, n = 28 \)
Figure 4.20  Histograms for coalescent simulations related to
D11S908: $\hat{b} = 0.5191, \hat{d} = 0.4809, \hat{\theta}_{SM,V} = 2.8380, n = 28$
Figure 4.21  Histograms for coalescent simulations related to D115908: $\mathbf{b} = 0.5191, \mathbf{d} = 0.4809, \mathbf{\hat{\theta}_{IAM,H}} = 2.5242, n = 28$
Figure 4.22  Histograms for coalescent simulations related to D11S908: $\hat{b} = 0.5191, \hat{d} = 0.4809, \hat{\theta}_{IAM,n_s} = 1.5169, n = 28$
the simulations generated by the $\theta_{IAM,na}$ estimate yielded the average number of alleles nearest the observed number of alleles; for D7S2434, the simulations using the $\theta_{IAM,H}$ estimate yielded the closest approximation to the observed number of alleles.

Figures 4.23 – 4.33 show a histogram of the sample and a histogram from the theoretical convolution of Poisson distributions for each of the loci, D1S2639, D2S169, D2S2290, D3S1317, D3S3635, D5S1364, D7S2434, D8S531, D9S117, D13S1269, and D14S1063.

<table>
<thead>
<tr>
<th>Chromosome</th>
<th>Number of Alleles in GDB Sample</th>
<th>Average Number of Alleles from Coalescent Simulations</th>
<th>$\theta_{IAM,H}$</th>
<th>$\theta_{IAM,na}$</th>
<th>$\theta_{SSM,H}$</th>
<th>$\theta_{SSM,v}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>D1S2639</td>
<td>5</td>
<td></td>
<td>2.451</td>
<td>3.353</td>
<td>3.306</td>
<td>2.878</td>
</tr>
<tr>
<td>D2S169</td>
<td>5</td>
<td></td>
<td>2.935</td>
<td>3.194</td>
<td>3.951</td>
<td>3.011</td>
</tr>
<tr>
<td>D2S2290</td>
<td>4</td>
<td></td>
<td>2.347</td>
<td>2.69</td>
<td>2.963</td>
<td>2.798</td>
</tr>
<tr>
<td>D3S1317</td>
<td>7</td>
<td></td>
<td>3.916</td>
<td>3.636</td>
<td>5.102</td>
<td>4.05</td>
</tr>
<tr>
<td>D3S3635</td>
<td>6</td>
<td></td>
<td>3.683</td>
<td>3.551</td>
<td>4.958</td>
<td>4.097</td>
</tr>
<tr>
<td>D5S1364</td>
<td>4</td>
<td></td>
<td>2.789</td>
<td>2.679</td>
<td>3.721</td>
<td>2.94</td>
</tr>
<tr>
<td>D7S2434</td>
<td>3</td>
<td></td>
<td>2.871</td>
<td>2.327</td>
<td>3.852</td>
<td>2.74</td>
</tr>
<tr>
<td>D8S531</td>
<td>7</td>
<td></td>
<td>3.645</td>
<td>3.725</td>
<td>5.57</td>
<td>4.475</td>
</tr>
<tr>
<td>D9S117</td>
<td>8</td>
<td></td>
<td>5.001</td>
<td>4.152</td>
<td>8.599</td>
<td>5.568</td>
</tr>
<tr>
<td>D11S908</td>
<td>6</td>
<td></td>
<td>3.989</td>
<td>3.276</td>
<td>6.117</td>
<td>4.083</td>
</tr>
<tr>
<td>D13S1269</td>
<td>6</td>
<td></td>
<td>3.863</td>
<td>4.986</td>
<td>5.154</td>
<td>3.974</td>
</tr>
<tr>
<td>D14S1063</td>
<td>4</td>
<td></td>
<td>2.906</td>
<td>2.664</td>
<td>3.326</td>
<td>2.846</td>
</tr>
</tbody>
</table>

Table 4.3 Comparison of average number of alleles generated in simulation with the observed number of observed from the GDB samples. The italicized number in each row is the estimate that most closely approximates the observed number of alleles for the given locus.
Sample plot for Dinucleotide locus D1S2639

Theoretical distribution under convolution of poisson's model

Figure 4.23  Histograms for D1S2639:
\[ \hat{b} = 0.5, \hat{d} = 0.5, \hat{\lambda} = 0.54, n_0 = 5, n = 27 \]
Figure 4.24  Histograms for D2S169:
\( \hat{b} = 0.6229, \hat{d} = 0.3771, \hat{\lambda} = 0.4844, n_a = 5, n = 56 \)
Figure 4.25 Histograms for D2S2290:
\[ \hat{b} = 0.0584, \hat{d} = 0.9416, \hat{\lambda} = 0.5304, n_a = 4, n = 28 \]
Sample plot for Dinucleotide locus D3S1317

Theoretical distribution under convolution of poisson's model

Figure 4.26 Histograms for D3S1317: 
\[ \hat{b} = 0.3540, \hat{d} = 0.6460, \hat{\lambda} = 0.9275, n_\text{a} = 7, n = 80 \]
Figure 4.27  Histograms for D3S3635:
\( \hat{b} = 0.9849, \hat{d} = 0.01514, \hat{\lambda} = 1.2744, n_a = 6, n = 56 \)
Figure 4.28  Histograms for D5S1364:
\[ \hat{b} = 0.8176, \hat{d} = 0.1824, \lambda = 0.4337, n_a = 4, n = 74 \]
Figure 4.29  Histograms for D7S2434:
\[ \dot{b} = 0.2204, \dot{d} = 0.7796, \lambda = 0.4664, n_a = 3, n = 28 \]
Figure 4.30 Histograms for D8S531:
\[ \hat{\theta} = 0.1426, \hat{d} = 0.8574, \hat{\lambda} = 1.4196, n_a = 7, n = 56 \]
Sample plot for Dinucleotide locus D9S117

Theoretical distribution under convolution of poisson's model

Figure 4.31  Histograms for D9S117:
\[ \hat{b} = 0.3209, \hat{d} = 0.6791, \hat{\lambda} = 2.1064, n_a = 8, n = 80 \]
Sample plot for Dinucleotide locus D13S1269

Theoretical distribution under convolution of poisson model

Figure 4.32 Histograms for D13S1269:
\[ \hat{b} = 0.6519, \hat{d} = 0.3481, \hat{\lambda} = 0.9979, n_a = 6, n = 54 \]
Sample plot for Dinucleotide locus D14S1063

Theoretical distribution under convolution of poisson's model

Figure 4.33 Histograms for D14S1063:
\( \hat{b} = 0.7087, \hat{d} = 0.2913, \hat{\lambda} = 0.4139, n_a = 4, n = 56 \)
Chapter 5

Estimation of Relative Mutation Rates of Microsatellite Loci

5.1 Principles of Estimation

Although we would like to be able to observe mutational events and thus calculate mutation rates for various microsatellite loci, indirect estimation may yield less biased estimates than those available from direct estimation due to mutational hot spots. As the mutation rate $\nu$ influences the distribution of observed allele sizes, a relationship between $\nu$ and the genetic variance at a locus may be established under the condition of mutation-drift balance within the population. The following method of estimating relative mutation rates was introduced by Zouros [29], and handled more rigorously by Chakraborty et al. [5].

In Section 4.6.1, genetic variance was given as $V/2 = \theta/2$, with $\theta = 4N\nu$, under the symmetrized single-step SMM at mutation-drift equilibrium, where $N$ is assumed to be of constant size throughout the generations. Note that the distribution of allele size changes is not at equilibrium due to the random walk nature of the mutations. However, the distribution of the difference of the sizes of two alleles randomly chosen from a population reaches mutation-drift equilibrium [6, 15].
Noting that

\[ V = E[(X_i - X_j)^2], \]

where \( X_i \) and \( X_j \) are the allele sizes on the \( i^{th} \) and \( j^{th} \) randomly chosen chromosomes at a given locus, \( i \neq j \), \( V \) reaches the equilibrium value

\[ V = 4N \nu \cdot Var(U_0), \]

where \( U_0 \) is the symmetrized distribution of the change of allele size due to mutation. If mutations are symmetric single-step, then \( Var(U_0) = 1 \) yielding

\[ V = 4N \nu = \theta, \]

as assumed in the estimate of \( \theta \) under the single-step SMM, previously stated.

From this we obtain an estimate of the mutation rate, up to a multiplicative constant

\[
\begin{align*}
\hat{V} &= 4N \hat{\nu} \\
\frac{\hat{V}}{2} &= 2N \hat{\nu} \\
&= \text{const} \cdot \hat{\nu}
\end{align*}
\]
given that population size is manageable. Here

\[ \hat{V}/2 = \text{sample variance of allele sizes for a locus, and} \]

\[ \hat{\nu} = \text{estimated mutation rate for a given locus.} \]

Between population variability is not a concern since we have only one population, Caucasians, in the GDB data we retained.
A log transform of the above gives us a log linear model for the relationship between \( \nu \) and \( V/2 \),

\[
\ln \left( \frac{V_i}{2} \right) = \ln(\nu) + \text{const} + \varepsilon_i,
\]

where

\( V_i/2 = \) the genetic variance associated with the \( i^{th} \) locus,

\( \text{const} = \) influence of population size,

\( \varepsilon_i = \) random error associated with the \( i^{th} \) locus, and

\( i \in 1 \ldots n \), where \( n \) is the number of loci observed.

Based on the theory that \( V = 4N\nu \) under the single-step SMM, and the assumption that the error term is approximately normally distributed with mean zero and variance \( \sigma^2 \), the above linear model should well approximate the mutation rate for a given motif type, up to an additive constant. Quantile-quantile plots of \( \ln \hat{V} \) for dinucleotides and tetranucleotides are shown in Figures 5.1 and 5.2. The q-q plot for the dinucleotides shows the distribution of \( \ln \hat{V} \) to be almost perfectly normal.

If the above assumptions are fulfilled, then

\[
\ln \nu = \frac{1}{n} \sum_{i=1}^{n} \ln \frac{\hat{V}_i}{2} - \ln 2N
\]

gives us an estimate of the mutation rate for a a given motif type.
Figure 5.1 Quantile-quantile plot of the log variance estimates for GDB dinucleotide data.

Figure 5.2 Quantile-quantile plot of the log variance estimates for GDB tetranucleotide data.
Note that the assumption of constant population size does not limit this analysis. The above linear model may also be applied to expanding populations, provided that the motif types of the loci studied are for data within the same populations [5].

We would like to be able to compare the rates of mutation of dinucleotide loci and tetranucleotide loci, collected from the same population. Using the theory developed above, we may estimate the relative magnitude of mutation rates for dinucleotides and tetranucleotides. To examine the ratio $\nu_{di}/\nu_{tetra}$, we may look at the exponent of the difference

$$\frac{1}{n_{di}} \sum_{i=1}^{n_{di}} \ln \frac{\hat{V}_i}{2} - \frac{1}{n_{tetra}} \sum_{i=1}^{n_{tetra}} \ln \frac{\hat{V}_i}{2}$$

since

$$\nu_{di}/\nu_{tetra} = \exp\{\ln\left(\frac{\nu_{di}}{\nu_{tetra}}\right)\} = \exp\{\ln(\nu_{di}) - \ln(\nu_{tetra})\}.$$ 

5.1.1 Relative Mutation Rates Based on GDB Data

The results of Chakraborty et al. [5] pertaining to four different data sets indicate that for non-disease related loci, mutation rates are inversely related to the size of the repeat motif. Two of the data sets were allele size distributions for di-, tri-, and tetranucleotide loci from different human populations. One of the remaining two sets was data on tetranucleotide loci for 13 human populations, and the other was a data set from GDB on di-, tri-, and tetranucleotide loci located on chromosome nineteen. Analyses of these data suggest that dinucleotides have a mutation rate
1.48 – 2.16 times higher than tetranucleotides, and non-disease related trinucleotide loci have a mutation rate 1.22 – 1.97 times higher rate than tetranucleotides, dependent on the data set. Disease-related trinucleotide loci, having observed allele sizes in the neutral range, showed a mutation rate 3.86 – 6.89 times higher than tetranucleotides – a rate exceeding dinucleotides as well. Chakraborty et al. [5] used a non-parametric Mann-Whitney test to detect location differences in the distribution of $\ln V$ for dinucleotides and tetranucleotides. Mann-Whitney is a rank test for two independent samples. For the Chromosome 19 GDB di- and tetranucleotide data alone, the Mann-Whitney test for values of $\ln V$ shows no significant difference in location between the distributions ($p = 0.24$).

Previously, researchers had concluded through direct studies that mutation rates for tetranucleotides were greater than those for dinucleotides [26, 28, 12]. In these studies, a small number of loci contributed a large portion of the observed in vivo mutations (that is, mutations within human cells) across the sample of loci, possibly biasing results for these studies due to one or two particular loci.

The GDB data from chromosome nineteen included in this study were related to the data used by Weber and Wong [26]. For the Weber and Wong study, removing the two loci that contained a disproportionate number of observed in vivo mutations yields results similar to the results in this study.

In the thesis study, we compared the relative mutation rates via the empirical distribution functions of $\ln \hat{V}$ for the di- and tetranucleotide data. First, a two-
sided Kolmogorov-Smirnov test for equality in distribution of the log variances of di- and tetranucleotides has a \( p \)-value of < 0.001. While the mutation rates were quite dispersed this indicates that

\[
\nu_{di} > \nu_{tetra}
\]

in general. Comparison of means of \( \ln \dot{V} \) yields

\[
\frac{\nu_{di}}{\nu_{tetra}} = 2.37,
\]

suggesting the dinucleotides have more than double the mutation rate of tetranucleotides, consistent with the estimates from Chakraborty et al. [5]. This tendency for estimates of mutation rate based on variance for the dinucleotide to be higher than those for tetranucleotides is apparent in Figure 5.3. As in other studies included in this thesis, there was not enough trinucleotide loci frequency information available in GDB to allow for inclusion of trinucleotides in this analysis.
Figure 5.3 Comparison plot of empirical cumulative distributions for di- and tetranucleotide GDB data on the first fourteen chromosomes.
Chapter 6

Conclusions

Through exploration of the large set of GDB microsatellite data for the first fourteen chromosomes in Caucasians, conclusions can be drawn about the mutational processes of microsatellites. This has been done in the three different studies conducted in this thesis.

First, an analysis of models of mutation revealed that both the IAM and the single-step SMM yield internally consistent estimates of a common parameter, \( \theta \), for dinucleotides. This consistency suggests that this method gives no power to differentiate between models for dinucleotide loci.

For tetrancleotides, a linear relationship exists between parameter estimates under both of the above models, and for both comparisons of biased and bias reduced estimates of \( \theta \); however, the slope of the least-squares line for the relationship of \( \theta_{IAM,n} \) versus \( \theta_{IAM,H} \) and \( \theta_{SMM,V} \) versus \( \theta_{SMM,H} \) is not near one. A model different from the equilibrium version of the IAM or the single-step SMM appears more appropriate.

In the second study we compared sample histograms to histograms generated under a convolution of Poissons model. Through conducting Kolmogorov-Smirnov tests comparing the theoretical convolution of Poisson distributions to the sample
distributions, we rejected the hypothesis that the sample distribution and the theoretical distribution were the same in four of twelve sample distributions studied. In two cases, the test statistic and the critical value were quite close, and thus we did not make inference about the feasibility of the hypothesis.

Also in the second study we compared the number of alleles observed in GDB samples with the number of alleles that appeared, on average, in 1000 simulations of allele size distributions. These simulations were generated using the estimates of backward mutation probability and the estimates of the parameter \( \theta \), under both the IAM and SMM. For ten of the twelve loci used in this study, the coalescent simulations that produced the average number of alleles nearest the observed number of alleles in the sample were those generated using estimates of \( \theta \) from the \( \theta_{SMM,H} \) estimator.

The third study, estimating relative mutation rates, yields results concurrent with those in the study by Chakraborty et al. [5]. Through a relationship between genetic variance and mutation rate, results were obtained that suggest that dinucleotides mutate at a rate more than twice that of tetranucleotides.

The results of these three studies suggest many things. First, they suggest that the tetranucleotide loci are not at mutation-drift equilibrium, as documented by the \( \approx 0.5 \) slope of their \( \theta - \theta \) plots. The dinucleotide loci are closer to equilibrium, based on the \( \approx 1.0 \) slope of their \( \theta - \theta \) plots. The estimates of the mutation rates
point in the same direction, suggesting that the tetranucleotide loci evolve more slowly, which is consistent with their larger distance from equilibrium.

Merit is given to the theory that the single-step SMM is feasible for modelling the mutations of dinucleotide loci. First, we found that the estimates of $\theta$ from the SMM for dinucleotides are fairly consistent. Also, the estimates of $\theta$ from the $\theta_{SMM,H}$ estimator generate coalescent samples with the number of alleles closest to the number observed for a sample at a given locus.

The relationships observed for tetranucleotides are consistent with the hypothesis that the population is expanding, but only if the SMM is accepted.

Indeed if the mutations of tetranucleotides in the Caucasoid population adhered to the assumptions of the IAM, and the population size increased from one equilibrium value, say $N_0$, to a new value, $N$, where $N \gg N_0$, then this increase in population size would logically result in an increase in the number of alleles produced at a locus, even with mutation rate remaining constant. The population of size $N$ would not be at mutation drift equilibrium, thus the new alleles created would not be removed by genetic drift. However, the new alleles created would be quite infrequent, causing the heterozygosity, $H$, to increase more slowly. This would cause estimates of $\theta$ based on number of alleles ($\hat{\theta}_{IAM,n_a}$) and based on heterozygosity ($\hat{\theta}_{IAM,H}$) to have an imbalance, such that $\hat{\theta}_{IAM,n_a} > \hat{\theta}_{IAM,H}$. We observed the opposite behavior, suggesting once again that the IAM is not compatible with our tetranucleotide loci.
Under the SMM, we see a different result. Numerical experiments presented by Kimmel, Chakraborty, and Jorde ([14], paper in progress) suggest that under the SMM, if the population size is expanding from some previous equilibrium value, the ratio of the expected value of the estimates of $\theta$ based on the variance $V(t)$ and on the heterozygosity $H(t)$, $\theta_{SMM,V(t)}$ and $\theta_{SMM,H(t)}$, will be less than 1 for many generations. Under stepwise, logistic, and exponential growth models, numerical simulations show rapid decline in the ratio $\theta_{SMM,V(t)}/\theta_{SMM,H(t)}$ in the first several generations of population size increase. Then, there is a slow increase in the ratio as the population size heads toward a new equilibrium value, not necessarily equal to 1. The numerical experiments for these growth models suggest that loci with high mutation rates, such as microsatellites, have a more extreme drop in the ratio $\theta_{SMM,V(t)}/\theta_{SMM,H(t)}$ in the first several generations of population expansion. The estimates of $\theta_{SMM,V}$ and $\theta_{SMM,H}$ for tetranucleotides in this thesis are consistent with the above theory of population expansion, with the estimated slope of the linear relationship between $\theta_{SMM,V}$ and $\theta_{SMM,H}$ near 0.5.
Bibliography


GDB Data Type: Polymorphisms.


