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.

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.

ProQuest Information and Learning
300 North Zeeb Road, Ann Arbor, MI 48106-1346 USA
800-521-0600

Genome Descent in Isolated Populations

Nicola H. Chapman

A dissertation submitted in partial fulfillment
of the requirements for the degree of

Doctor of Philosophy

University of Washington

2001

Program Authorized to Offer Degree: Biostatistics

UMI Number: 3013940

UMI Microform 3013940

Bell & Howell Information and Learning Company
300 North Zeeb Road
P.O. Box 1346
Ann Arbor, MI 48106-1346

In presenting this dissertation in partial fulfillment of the requirements for the Doctorial degree at the University of Washington, I agree that the Library shall make its copies freely available for inspection. I further agree that extensive copying of this thesis is allowable only for scholarly purposes, consistent with "fair use" as prescribed in the U.S. Copyright Law. Requests for copying or reproduction of this dissertation may be referred to University Microfilms, 1490 Eisenhower Place, P.O. Box 975, Ann Arbor, MI 48106, to whom the author has granted "the right to reproduce and sell (a) copies of the manuscript in microform and/or (b) printed copies of the manuscript made from microform."

Signature Nicola Chapman

Date June 1, 2001

University of Washington

Graduate School

This is to certify that I have examined this copy of a doctoral dissertation by

Nicola H. Chapman

and have found that it is complete and satisfactory in all respects,
and that any and all revisions required by the final
examining committee have been made.

Chair of Supervisory Committee:

Elizabeth Thompson

Elizabeth A. Thompson

Reading Committee:

Elizabeth Thompson

Elizabeth A. Thompson

Ellen M. Wijsman

Ellen M. Wijsman

Joseph Felsenstein

Joseph Felsenstein

Date: June 1, 2001

University of Washington

Abstract

Genome Descent in Isolated Populations

by Nicola H. Chapman

Chair of Supervisory Committee

Professor Elizabeth A. Thompson
Statistics

An isolated population is one that is descended from a small group of individuals (founders), and in which population growth is due almost exclusively to births within the population, rather than immigration from outside. Interest in the genetics of isolated populations has recently been revived among medical geneticists, because it is hoped that diseases for which there are several susceptibility loci in large outbred populations may be more homogeneous in small isolated populations. In addition, it has been suggested that small recently founded populations may exhibit linkage disequilibrium over longer genetic distances than large outbred populations.

In this dissertation, I study the inheritance of segments of ancestral chromosomes. This is achieved by the study of junctions - points on the chromosome where segments from different ancestral chromosomes meet. I consider a discrete generation random mating population, and assume that population sizes over time are known. Two fundamental questions are addressed. First, how large are pieces of intact ancestral chromosome in chromosomes sampled from the population? The answer to this question provides information about the extent of linkage disequilibrium in the population. Second, how long are the tracts of chromosome which are shared identical-by-descent between randomly sampled chromosomes from the population? I also study how the age of the population, founding population size, growth patterns, and population subdivision affect both segment length, and the length

of IBD tracts between randomly sampled chromosomes. I show that these factors have substantial effects in very small populations, but are less important in larger populations.

The theoretical work about junctions in random mating populations is extended to the case where a pedigree is available. I introduce an approach to the estimation of fine-scale genetic maps, using data consisting of chromosomes sampled from an isolated population, and knowledge of the pedigree relating those chromosomes. The approach is demonstrated using a pedigree relating 27,163 Hutterites. I show that this approach has potential to produce more precise maps than currently exist, and that the Hutterite population is ideal for this application.

TABLE OF CONTENTS

List of Figures vi

List of Tables xii

Chapter 1: Introduction 1

1.1 Junctions 2

1.2 Simulations on the Hutterite Ancestry 5

1.3 Application of Junction Theory to Map Estimation Using Hutterite Chromosomes 6

Chapter 2: The Effects of Population Size and Growth on Junction Number and Sharing 9

2.1 Number of Junctions per Morgan 10

2.1.1 Expectation 10

Constant Population Size 11

Growing Populations 13

Example populations with different types of growth 14

2.1.2 Variance 18

Covariance 21

Constant Population Size 24

Example populations with different types of growth 26

2.2 Number of junctions shared per Morgan 28

2.2.1 Expected number shared between two chromosomes 28

Calculation of $Pr(Z_t = 2 \mid Y_{j+1} = 1)$ 29

Constant Population Size 31

Growing Populations 34
Example populations with different types of growth 34
2.3 Discussion 39

Chapter 3: The Effects of Random Subdivision and Non-Random Mating on Junction Number and Sharing 41

3.1 Population subdivision 41
3.1.1 Expected number of junctions per Morgan 41
Constant Size Population 42
Example of growing populations with repeated subdivision 43
3.1.2 Variance of number of junctions per Morgan 47
Example of growing populations with repeated subdivision 47
3.1.3 Expected number of junctions shared per Morgan 49
Example of growing populations with repeated subdivision 50
3.2 Regular Mating Systems 52
3.2.1 Sib mating 52
3.2.2 Double-first-cousin mating 55
3.2.3 First-cousin mating 58
E[J_t] diverges 62
3.3 Discussion 64

Chapter 4: The Effects of Population Size, Growth and Subdivision on the Lengths of IBD Regions 66

4.1 Modelling the length of an IBD tract 68
4.1.1 Model A: Independence of Junction Types and IID Exponential Segments 68
Assessing model fit 69
4.1.2 Model B: 1st order Markov dependence of junction types and I.I.D. exponential segments 73

Assessing model fit 77
4.1.3 Model C: 1st order Markov dependence of junction types, IBD and non-IBD segments modelled separately 83
Assessing model fit 84
Relationship to work of Stam [27] 85
Modelling the variance of the length of an IBD tract 85
4.2 Application to growing populations without subdivision 87
4.3 Application to growing populations with subdivision 91
4.3.1 Extension of Stam's work to subdivided populations 91
4.3.2 Example of the effects of population subdivision 92
4.4 Discussion 101

Chapter 5: Simulating Chromosome Transmissions in Random Mating Populations and on the Hutterite Ancestry 102

5.1 Data Structures 102
5.1.1 Individuals 102
5.1.2 Segments and Chromosomes 103
5.2 Algorithms 103
5.2.1 Gamete Production 104
5.2.2 Random Mating Populations 104
5.2.3 Pedigree Based Simulation 105
5.2.4 Computational Demands 105
5.3 Simulations on the Hutterite Ancestry 106
5.3.1 Hutterite History, Condensed 106
5.3.2 The Hutterite Pedigree 107
5.3.3 Approximating the Hutterite population by a random mating population 108
5.3.4 Number of junctions in a randomly selected chromosome 113
Theoretical predictions based on Hutterite population sizes 113

Simulation results 115
5.3.5 Lengths of IBD tracts in pairs of Hutterite chromosomes 118
Theoretical predictions based on Hutterite population sizes 118
Simulation results 119
5.4 Discussion 124

Chapter 6: Application of Junction Theory to Map Estimation 126
6.1 Moments of the number of junctions per Morgan, conditional on pedigree structure 126
6.1.1 Mean 126
6.1.2 Variance 128
6.1.3 Covariance 130
Calculation of $E[S_{c,d}]$ 134
6.2 Quasi-likelihood modelling of junction counts 136
6.2.1 Independence 137
6.2.2 Covariance approximated by $E[S]$ 139
6.2.3 Effects of sample composition on the variance of the estimate 140
6.3 Detection and resolution of junctions using continuous IBD data 143
6.3.1 Characteristics of undetected or unresolved junctions 146
6.3.2 Effects of sample composition on detection and resolution 148
6.4 Discussion 151

Chapter 7: Discussion 152
7.1 Implications for studies of disequilibrium in isolated populations 153
7.2 Implications for estimation of genetic distance in isolated populations 154

Bibliography 156

Appendix A: Moments of Non-Identity Proportions in a Random Mating Population 159

Appendix B: Moments of the Number of Junctions formed in Meioses from a Given Generation of a Random Mating Population 164

LIST OF FIGURES

1.1 Examples of chromosomes in isolated populations. Different shadings represent different ancestral types. 3

1.2 Formation and transmission of junctions. 4

2.1 Expected number of junctions per Morgan, for a population of constant size N. 12

2.2 Expected number of junctions per Morgan, as a function of the inbreeding coefficient. 12

2.3 Expected number of junctions per Morgan, for populations in which $N_0 = 20$, and $m = 0$, 0.1, 0.2 or 1.0. 15

2.4 Population sizes for linear and exponential growth, $N_0 = 20$ and $N_{100} = 2020$. 16

2.5 Expected number of junctions per Morgan as function of generation, for populations growing exponentially and linearly. Founder population sizes are (a) 20, (b) 100, and (c) 500. 17

2.6 Empirical distributions of $J_{50} \mid N$ and $J_{150} \mid N$, for $N = 20$ and $N = 50$. . . 19

2.7 Variance of the number of junctions in a randomly selected chromosome from a population of constant size $N = 20$ or $N = 50$; estimated by (i) simulation, (ii) Equation 2.27, and (iii) the Poisson approximation. 25

2.8 Variance of the number of junctions in early generations from a population of constant size $N = 50$; estimated by (i) simulation, (ii) Equation 2.27, and (iii) the Poisson approximation. 26

2.9 Distributions (estimated from 10,000 simulations) of the number of junctions in a randomly selected chromosome from generation 100, for populations growing either linearly or exponential with $N_0 = 20$. 27

2.10 Expected number of junctions in a randomly selected chromosome from a population of constant size $N = 20$ or $N = 50$; (i) total, (ii) shared IBD with its homologue, and (iii) not shared IBD with its homologue. 33

2.11 Expected number of junctions in a randomly selected chromosome from a linearly growing population with $N_0 = 20$ and $m = 0.10$, $m = 0.25$ or $m = 1.00$; (i) total, (ii) shared IBD with its homologue, and (iii) not shared IBD with its homologue. 35

2.12 Expected number of junctions in a randomly selected chromosome from a population growing either linearly or exponentially from $N_0 = 20$; (i) total, (ii) shared IBD with its homologue, and (iii) not shared IBD with its homologue. 36

2.13 Expected number of junctions in a randomly selected chromosome from a population growing either linearly or exponentially from $N_0 = 100$; (i) total, (ii) shared IBD with its homologue, and (iii) not shared IBD with its homologue. 37

2.14 Expected number of junctions in a randomly selected chromosome from a population growing either linearly or exponentially from $N_0 = 500$; (i) total, (ii) shared IBD with its homologue, and (iii) not shared IBD with its homologue. 38

3.1 Expected number of junctions in a population of size (a) N=20 or (b) N=50 divided into 2 subpopulations at varying times. 44

3.2 Expected number of junctions per Morgan as function of generation, for exponentially growing populations with varying degrees of subdivision. Founder population sizes are (a) 20, (b) 100, and (c) 500. 46

3.3 Distributions (estimated from 10,000 simulations) of the number of junctions in a randomly selected chromosome from generation 100, for each of 4 growth and subdivision scenarios with $N_0 = 20$. 48

3.4 Expected number of junctions shared per Morgan as function of generation, for exponentially growing populations with varying degrees of subdivision. Founder population sizes are (a) 20, (b) 100, and (c) 500. 51
3.5 Repeated sib-mating. 53
3.6 Expected number of junctions per Morgan in a chromosome sampled from generation t, repeated sib mating. 54
3.7 Repeated double-first-cousin mating. 56
3.8 Expected number of junctions per Morgan in a chromosome sampled from generation t, repeated double first cousin mating. 57
3.9 Repeated first-cousin mating. 59
3.10 Expected number of junctions per Morgan in a chromosome sampled from generation t, repeated first cousin mating. 61

4.1 Two chromosomes sampled from a population, some time after founding. Different shades represent different ancestral chromosomes. 67
4.2 Mean length of an IBD tract as estimated by simulation and by models A, B and C, for populations of size (a) N=20, and (b) N=100. 70
4.3 Mean number K of type S junctions in an IBD tract, estimated by simulation and by models A, B&C, for populations of size (a) N=20, and (b) N=100. . 72
4.4 Estimates of equilibrium probabilities π_S and π_T based on the theoretical approximation and 10,000 simulations, for populations of constant size (a) N=20 and (b) N=100. 76
4.5 Distribution of segment length as a function of segment type, estimated by 100,000 simulations, for a population of size N=20 at generations (a) 5, (b) 10, and (c) 20. 82
4.6 Variance of the length of an IBD tract, as estimated directly by simulation, or as the square of the simulated mean, for populations of constant size (a) $N = 20$ and (b) $N = 100$. 86
4.7 Population sizes over time for linear and exponentially growing populations. . 88

4.8 Expected length of an IBD tract, for populations expanding 100-fold over 100 generations, either linearly or exponentially, with (a) $N_0 = 20$, (b) $N_0 = 100$, and (c) $N_0 = 500$. 89

4.9 Estimated density, based on 1000 simulations, of the length of an IBD tract at generation 100, for exponentially and linearly growing populations with $N_0 = 20$ and $N_{100} = 2020$. 90

4.10 Expected length of an IBD tract, between two chromosomes from within the same sub-population. Different lines within a plot correspond to populations with different amounts of subdivision. The lines from the top go from most subdivided to least subdivided, with the bottom line being the non-subdivided population. The three plots correspond to different starting sizes of (a) 20, (b) 100, and (c) 500. All populations are growing at a constant exponential rate of 4.72% per generation. 93

4.11 Expected length of an IBD tract, for an exponentially growing population with $N_0 = 20$ and $N_{100} = 2020$, and different levels of subdivision. The populations bifurcate whenever (a) $N = 40$, (b) $N = 80$, or (c) $N = 160$. . . 95

4.12 Expected length of an IBD tract, for an exponentially growing population with $N_0 = 100$ and $N_{100} = 10,100$, and different levels of subdivision. The populations bifurcate whenever (a) $N = 200$, (b) $N = 400$, or (c) $N = 800$. . 97

4.13 Expected length of an IBD tract, for an exponentially growing population with $N_0 = 500$ and $N_{500} = 50,500$, and different levels of subdivision. The populations bifurcate whenever (a) $N = 1000$, (b) $N = 2000$, or (c) $N = 4000$. 99

4.14 Estimated density, based on 1000 simulations, of the length of an IBD tract at generation 100, for exponentially growing populations with $N_0 = 20$, $N_{100} = 2020$, and varying levels of subdivision. 100

5.1 Schematic of chromosome representation. 103

5.2 Schematic of gamete production. 104

5.3 Mean number of junctions in a chromosome of length one Morgan, in both maternal and paternal chromosomes of selected Hutterites. Vertical lines represent theoretical means based on models random-25 (left lines) and random-20 (right lines). 116

5.4 Distributions of the number of junctions in one Morgan, estimated by 5,000 simulations each for models random-20 and random-25. Short vertical lines above the x-axis mark the observed mean number of junctions in 18 chromosomes from the appropriate leut of the Hutterite population, based on 100,000 simulations on the Hutterite pedigree. 117

5.5 Mean length of a random IBD tract between the paternal chromosomes of selected Hutterites, within the same leut. Vertical lines represent theoretical means based on models random-20 (left lines) and random-25 (right lines). . 120

5.6 Mean length of a random IBD tract between the paternal chromosomes of selected Hutterites, between different leut. Vertical lines represent theoretical means based on models random-20 (left lines) and random-25 (right lines). . 122

5.7 Distribution of the length of a random IBD tract between chromosomes within leut, estimated based on 5,000 simulations each of models random-20 and random-25. Short vertical lines above the x-axis mark the observed mean length of an IBD tract in 36 chromosome pairs from the appropriate leut of the Hutterite population, based on 100,000 simulations on the Hutterite pedigree. 123

5.8 Distribution of the length of a random IBD tract between chromosomes chosen from different leut, estimated based on 5,000 simulations each of models random-20 and random-25. Short vertical lines above the x-axis mark the observed mean length of an IBD tract in 81 chromosome pairs from the appropriate leut-pair of the Hutterite population, based on 100,000 simulations on the Hutterite pedigree. 125

6.1 Estimated variance based on 10,000 simulations vs. theoretical mean, for 200 chromosomes from 100 randomly sampled Hutterites. 130
6.2 Estimated covariance based on 10,000 simulations vs. expected number of junctions shared, for all distinct pairs of 200 chromosomes from 100 randomly sampled Hutterites. 136
6.3 Simulation distributions of estimates of interval length based on the independence assumption, for true values of $d = 5, 2, 1$ and 0.5 cM. Data used were 200 chromosomes from 100 randomly sampled Hutterites. 138
6.4 Simulation distributions of estimates of interval length using $Cov(J_c, J_d) = \mathrm{E}[S_{c,d}]$, for true values of $d = 5, 2, 1$ and 0.5 cM. Data used were 200 chromosomes from 100 randomly sampled Hutterites. 141
6.5 Four kinds of chromosome with respect to junction depicted. Different shadings represent different ancestral types. 144
6.6 Detection and resolution of a junction using continuous IBD data. On the left, different shadings represent different ancestral types, and on the right, white represents IBD and black non-IBD. 144

A.1 Two locus gene non-identity measures. 160

LIST OF TABLES

2.1 Expected number of junctions per Morgan in a chromosome taken from generation 100, for exponential and linear growth. 16

2.2 Variance of the number of junctions in a chromosome randomly selected from generation 100, based on 10,000 simulations, the Poisson approximation, or Equation 2.27. 28

3.1 Expected number of junctions per Morgan in a chromosome taken from generation 100, for four subdivision scenarios in populations growing exponentially at 4.7% per generation. 47

3.2 Variance of the number of junctions in a chromosome randomly selected from generation 100, based on 10,000 simulations, the Poisson approximation, or Equation 2.27. 49

4.1 Empirical distribution of K compared to that predicted under Model A, for N=20, t=5, 10 and 20. 71

4.2 Empirical distribution of K compared to that predicted under Model A, for N=100, t=5, 10 and 20. 73

4.3 Estimates of the transition matrix parameters, based on 100,000 simulations, for $N = 20$ and $t = 5$, 10 or 20. 78

4.4 Empirical distribution of K compared to that predicted under Models B&C, for N=20, t=5, 10 and 20. 78

4.5 Estimates of the transition matrix parameters, based on 100,000 simulations, for $N = 100$ and $t = 5$, 10 or 20. 79

4.6 Empirical distribution of K compared to that predicted under Models B&C, for N=100, t=5, 10 and 20. 79

4.7 Mean and median lengths of different types of segments, based on 100,000 simulations, for a population of size $N = 20$. 80

4.8 Mean and median lengths of different types of segments, based on 10,000 simulations, for a population of size $N = 100$. 81

4.9 Expected length of an IBD tract in two randomly chosen chromosomes from generation 100. 90

4.10 Expected length of an IBD tract for chromosomes chosen within or between sub-populations at generation 100, for an exponentially growing population with $N_0 = 20$ and $N_{100} = 2020$. 96

4.11 Expected length of an IBD tract for chromosomes chosen within or between sub-populations at generation 100, for an exponentially growing population with $N_0 = 100$ and $N_{100} = 10,100$. 96

4.12 Expected length of an IBD tract for chromosomes chosen within or between sub-populations at generation 100, for an exponentially growing population with $N_0 = 500$ and $N_{100} = 50,500$. 98

5.1 Examples of computational demands of simulation of random mating populations. 106

5.2 Number of individuals in 20 year birth cohorts, classified according to founder status and known leut information. 109

5.3 Population sizes used for a discrete generation, subdivided, randomly mating population approximating the Hutterites, based on 20 year birth cohorts. . . 110

5.4 Number of individuals in 25 year birth cohorts, classified according to founder status and known leut information. 111

5.5 Population sizes used for a discrete generation, subdivided, randomly mating population approximating the Hutterites, based on 25 year birth cohorts. . . 112

5.6 Expected number (standard deviation) of junctions per Morgan in a chromosome randomly selected from the most recent generation of L-leut, S-leut and D-leut, based on models random-20 and random-25. 113
5.7 Expected length in Morgans of an IBD tract between two randomly chosen chromosomes from the most recent generation of Hutterites, based on models random-20 and random-25. 118

6.1 Results of 1,000 simulations estimating d assuming independence and using 200 chromosomes from 100 randomly sampled Hutterites. 139
6.2 Results of 1,000 simulations estimating d using $Cov(J_c, J_d) = \mathrm{E}[S_{c,d}]$ and 200 chromosomes from 100 randomly sampled Hutterites. 141
6.3 Simulation results comparing estimation of $\hat{d}$, using different data sets. . . . 143
6.4 Simulation results: average number of junctions existing, average proportion detected, average proportion resolved in 200 chromosomes from 100 randomly sampled Hutterites. 145
6.5 Simulation results: average number of junctions existing, average proportion of junctions detected and average proportion of detected junctions that are resolved, by formation cohort, estimated by 500 simulations over an interval of 50 cM. (200 chromosomes from 100 randomly sampled Hutterites). 147
6.6 Simulation results: average number of junctions existing, average proportion detected, average proportion resolved, over 1,000 simulations of an interval of length 5 cM. 148
6.7 Simulation results: average proportion of junctions detected and average proportion of detected junctions that are resolved, by formation cohort and data set, estimated by 500 simulations over an interval of 50 cM. 150

A.1 Possible configurations of a, a', b, and b'. 162

ACKNOWLEDGMENTS

I would like to thank Elizabeth Thompson for her guidance and support, and in particular for her generosity with her valuable time. Thanks are also due to Ellen Wijsman and Joe Felsenstein, for reading this dissertation and providing valuable comments and suggestions. I am grateful to Joyce Crumley, T. Mary Fujiwara, and Kenneth Morgan for collecting, maintaining, and allowing me to use the Hutterite pedigree. Thanks also to J.C. Loredo-Osti for maintaining the computer on which the Hutterite analyses were run, and to Stan Chapman for helpful discussions about mathematical issues. Financial support for some of this work came from the Burroughs Welcome Fund (BWF) for the Program in Mathematical and Molecular Biology (PMMB). This work could not have been completed without the constant loving support of Jeff Noyle.

DEDICATION

To my family, and Tadders in particular, with all my love.

Chapter 1

INTRODUCTION

An isolated population is a modern population that is descended from a small group of individuals (founders) who lived some time ago. Population growth is achieved mostly by births within the population rather than immigration from outside. Interest in the genetics of isolated populations has recently been revived among human geneticists, because of suggestions that such populations may be useful for disequilibrium based mapping of susceptibility loci for complex disease. In particular, it is hoped that diseases for which there are several susceptibility loci in large outbred populations may be more homogeneous in small isolated populations. In addition, small recently founded populations may exhibit linkage disequilibrium over longer genetic distances than large outbred populations (see [5], [21]).

Isolated populations are fundamentally different from the large outbred populations that are usually assumed in the theoretical study of linkage disequilibrium. Disequilibrium can be affected by small population size, which results in genetic drift, and therefore by patterns of population growth. There is often substantial structure within isolated populations, which may have an effect on linkage disequilibrium. Examples include large scale population subdivision, often further division within subpopulations (which may be random or family based), and non-random mating within subpopulations. For a brief survey of the variety of histories and structures seen in human populations, we refer the reader to [4].

The utility of a population for a study based on disequilibrium testing may depend on its history. A recent paper [22] presented observed disequilibria in two regions of the genome for a wide variety of human populations. Their results showed that in general, levels of disequilibrium in isolated populations were only slightly higher than in outbred populations. While this result is somewhat disappointing, the pairs of loci considered in

this study were very tightly linked (< 0.2 cM apart), and so this result may simply reflect the large number of generations required to break down such associations. In a discussion of this paper [20], it was noted that the results are specific to the two chromosomal regions considered, and therefore may not be reflective of the situation in the rest of the genome. Also, there was one isolate in which disequilibrium was significantly higher than in the larger populations, demonstrating that the unique evolutionary history of each isolate must be considered.

These observations suggest that there is a need for systematic empirical study of disequilibrium across the entire genome, both within isolates and within larger outbred populations, to identify populations in which disequilibrium stretches over distances appropriate for susceptibility locus detection [20]. In this dissertation, we study theoretically the effect of several aspects of population history and structure on chromosomes from isolated populations. This work is therefore intended to complement empirical studies of disequilibrium in human populations.

1.1 Junctions

Ideally, we would like to study how the inheritance of contiguous ancestral chromosome segments is affected by population history and structure. Study of the ancestry of segments is a difficult problem, as it requires that we account for the possibility of recombination. The approach we take in this dissertation is to study the formation and transmission of the ends of contiguous ancestral chromosome segments. These ends are called *junctions*, since they are the point where chromosome segments from two distinct ancestors meet. Junctions are easier to study than segments since each is only a single point on the chromosome.

Figure 1.1 shows examples of two chromosomes that might have been sampled from a modern isolated population. Different shadings represent different ancestral types. The top chromosome contains two junctions, and the chromosome is therefore made up of three segments. The bottom chromosome contains eight junctions, and is made up of nine segments. A quantity of interest is the average length of contiguous ancestral segments remaining in the generation under study. If the chromosomes have broken into many short pieces rela-

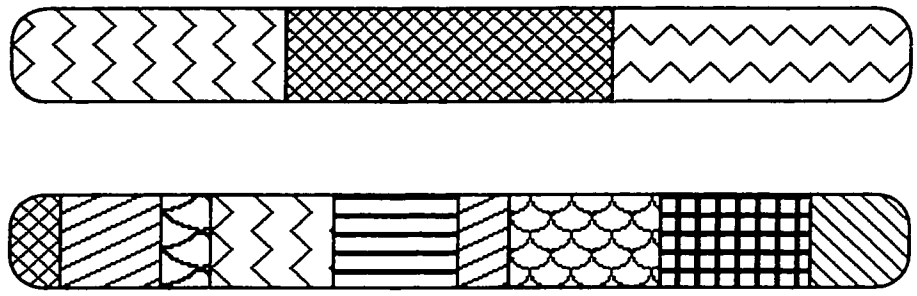


Figure 1.1: Examples of chromosomes in isolated populations. Different shadings represent different ancestral types.

tive to the founder population, disequilibrium due to founder effect will stretch over only short distances. Conversely, if the chromosomes are composed of a small number of large pieces, relative to the founder generation, disequilibrium will stretch over longer distances. If there are J junctions in a chromosome, there are $J+1$ ancestral segments, and by Jensen's inequality,

$$\mathrm{E}[\text{length of a segment}] = \mathrm{E}\left[\frac{1}{J+1}\right] \geq \frac{1}{\mathrm{E}[J]+1} \,.$$

Thus by obtaining the expected number of junctions in a length of chromosome, we obtain the expected number of contiguous segments, and therefore a lower bound on their expected length.

Junctions were initially defined in the context of plant and animal breeding by Fisher [9], who also described their formation and transmission. Fisher [9] considered the expected number of junctions formed and surviving in a regular sib-mating system. A junction is formed when a recombination occurs between two chromosomes, at a point where they are not descendants of the same ancestral chromosome. That is, the chromosomes are not *identical by descent* (IBD) at that point. Once a junction is formed, it is transmitted as is any other gene (according to the laws of Mendelian inheritance). Figure 1.2 illustrates the formation and transmission of junctions, in three meioses. The ancestral origin of each chromosome is indicated by its shading. In meiosis one, two recombinations occur, both in regions where the chromosomes are not IBD, so both are visible as junctions in the resulting gamete. In meiosis two, no recombinations occur, so no junctions are formed. In meiosis

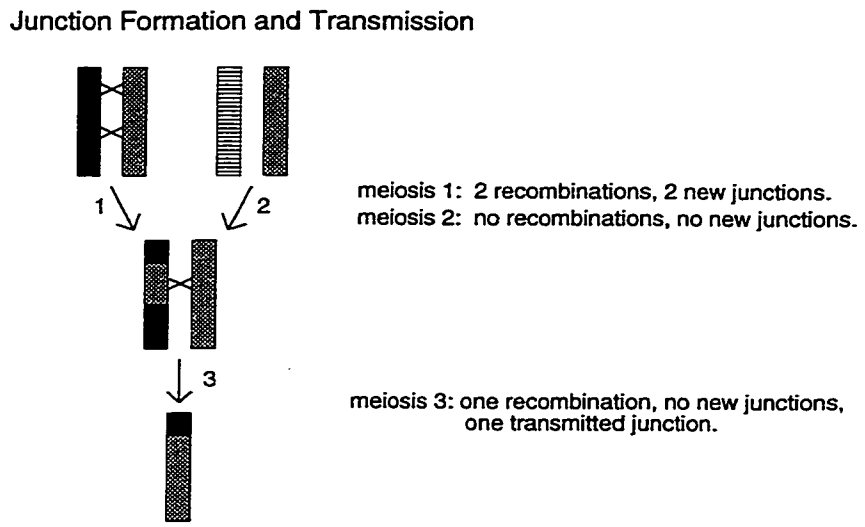


Figure 1.2: Formation and transmission of junctions.

three, one recombination occurs, but it is in a region where the chromosomes are IBD, so no junction is formed. However, one of the junctions formed in meiosis one is transmitted in meiosis three. These examples illustrate two important features of junctions. First, they are formed by recombination events between non-IBD chromosomes, and second, once formed, they are transmitted like any other Mendelian gene. It is important to realize that junctions, since they require non-IBD to be formed, are always relative to some ancestral population. In this work, the founding generation is assumed to consist of non-inbred, unrelated individuals, and we consider junctions relative to this population.

Fisher also made a distinction between *external* and *internal* junctions. A junction is internal if the chromosome is either IBD on both sides with its homologue, or non-IBD on both sides with its homologue. An external junction is one where the chromosomes are IBD on one side and non-IBD on the other. External junctions therefore mark the ends of IBD and non-IBD tracts between pairs of chromosomes. Knowing the number of external junctions between a pair of chromosomes therefore gives the number of sections into which the total length IBD (or non-IBD) is divided. A series of papers by Fisher ([10], [11]), Bennett [1] and Gale [13] calculated the expected number of external junctions in selfing, sib-mating, and parent-offspring mating. Bennett [2] and Fisher [10] both considered the

asymptotic distribution of the total length of genome non-IBD, using these results. Using a similar approach, Franklin [12] calculated the variance of the total length of genome non-IBD for any generation in selfing, sib-mating and parent-offspring systems. Stam [27] extended Fisher's ideas to a random mating population of constant size, and obtained the expected number of external junctions in an individual from any given generation. He also considered the asymptotic distribution of the total length of genome IBD in such a population.

In this dissertation, we apply these ideas to random mating populations whose generation sizes over time are known. In Chapter 2, we calculate the expected number of junctions per Morgan in a chromosome randomly selected from the population at a given time, and develop an approximation to the variance of this quantity. We explore the effects that different types of population growth can have on the expected number of junctions per Morgan, and we also consider the expected number of junctions shared between two chromosomes sampled from the same generation. In Chapter 3, we study the effects of random subdivision on the expected number of junctions per Morgan. We then present some new results regarding the expected number of junctions per Morgan in double-first-cousin and first-cousin mating systems. In Chapter 4, we discuss an approach to modelling the length of IBD tracts between randomly sampled chromosomes. By extending Stam's results to apply to subdivided populations, we explore the effects of population subdivision on the mean length of IBD tracts.

1.2 Simulations on the Hutterite Ancestry

As an example of an isolated human population, we consider the North American Hutterites. The Hutterites are a German speaking religious group who live on communally owned farming colonies, in the prairie provinces and states of North America. In 1994, there were approximately 35,000 Hutterites living in 382 colonies [16], all of whom were descended from 443 individuals who immigrated to the United States from the Ukraine between 1874 and 1879 [8]. The Hutterites' distinctive religious practices include communal living, adult baptism and pacifism, and they believe in the separation of church and state [16]. These beliefs have resulted in almost complete isolation from the secular world, and the extremely

rapid growth of the Hutterites since 1880 has been achieved entirely by a high birth rate. For example, the average completed family size for Hutterite women aged 45-54 in 1950 was 10.6 children [8].

The Hutterite population is unusually rich in structure. Within the population there exist three distinct sub-populations, which are known as leut. The split into leut occurred around the time of the migration to North America, and since the early 1900s there has been almost no inter-leut marriage. The population is also structured within each of the leut. Hutterite colonies divide into two daughter colonies when they become large enough and have saved enough money to purchase more land. Marriages in which two or more siblings marry individuals who are themselves siblings are common, and a woman always moves to her husbands' colony when she marries.

Through collaboration with Dr. Ken Morgan, of the University of Montreal, we have access to a large pedigree which traces the ancestry of the cohort of Hutterites alive in 1981 back to ancestors who were born in the 1700s. By simulating chromosome transmissions on this ancestry, we can compare the number of junctions observed in Hutterite chromosomes to the number expected based on the theory developed in Chapters 2 and 3. This allows us to consider the effects on expected junction number of structure which is too complicated to model theoretically.

In Chapter 5, we discuss some of the computational issues associated with simulating chromosome transmissions in random mating populations and the Hutterite pedigree. We first describe some of the data structures and algorithms used, and then present results of some simulations on the Hutterite pedigree. These results are compared with theoretical expectations for random mating populations with historical population sizes similar to the Hutterites.

1.3 Application of Junction Theory to Map Estimation Using Hutterite Chromosomes

Multipoint methods for linkage analysis and disequilibrium analysis require a map of polymorphic markers, where the order of markers and the recombination fractions between them

are assumed known. Marker maps are generated by typing three generation families. Information on the grandparents sets phase in the parents, and recombinants are counted in the children. The recombination fraction is then estimated by the the number of recombinants observed, divided by the total number of meioses. The variance of this estimate is then given by $\frac{\theta(1-\theta)}{n}$, where θ is the true recombination fraction, and n is the total number of meioses observed. Maps in use today were estimated using 8 of the CEPH reference families with a total of about 200 meioses. This number of meioses is too few to estimate precisely the distance between closely linked markers. For example, if the true recombination fraction between two markers is 0.05, the standard deviation of the estimate based on 200 meioses is 0.015. More accurate maps are desirable, since it has been shown [7] that misspecification of map distances can lead to both a reduction in power to detect linkage, and an increase in Type I error in multipoint linkage analysis. While the effect is small in the case of a single-gene disease, it has been suggested [7] that in the case of a complex trait, the problem could be more serious. Although the recent completion of the Human Genome Project may provide information on the order of markers on a chromosome, the fact that the estimated number of centiMorgans per megabase in the sex-averaged map varies from 0 to 6 across the genome [29] shows that good estimates of genetic distance cannot be obtained from physical distance.

We therefore consider the estimation of map distance using chromosomes sampled from the Hutterites, and knowledge of the pedigree relating them. Recombination events in the history of the population are visible in chromosomes sampled from the modern population as junctions, since junctions are recombination events that occur in a non-IBD region of the chromosome. The junctions represented in the modern population therefore reflect recombination events that occurred in a large number of ancestral meioses. Estimation of genetic distance based on junctions can be thought of as considering more meioses, and may result in more precise estimates than traditional methods. Since on average half the meioses making up the ancestry of a particular chromosome will be from males and half will be from females, a map obtained in this way will be a sex-averaged map.

In Chapter 6, we extend the mean, variance and covariance calculations of Chapter 2 so that they may be applied to chromosomes in individuals whose pedigrees are known. We

then describe a quasi-likelihood model for estimating the genetic distance spanned by an interval, assuming that the data consist of Hutterite chromosomes in which all junctions in the interval can be counted. We explore the performance of the method by simulating data on the Hutterite pedigree, and finally we consider how one might detect and assign junctions to chromosomes based on continuous IBD information, which is a first step towards making the method applicable to real data.

Chapter 2

THE EFFECTS OF POPULATION SIZE AND GROWTH ON JUNCTION NUMBER AND SHARING

In this chapter, we consider junction formation and transmission in a random mating monoecious population, with discrete generations. The founder generation consists of N_0 unrelated, diploid individuals, whose genetic material is carried on a single pair of homologous chromosomes. Chromosomes in this group are assumed to be of distinct ancestral origin. The N_t individuals in the tth generation give rise to the N_{t+1} individuals in the $t+1$th generation according to the following algorithm. For an individual in generation $t+1$, two parents are chosen randomly, with replacement, from the individuals in generation t. A gamete is formed by each parent, and these are united to make up the genotype of the offspring. This process is repeated for every individual in generation $t+1$. We assume that during gamete formation, recombination events along the chromosome happen according to a Poisson process, which necessarily has rate one per Morgan. This implies that the number of recombinations in a chromosome of length L has a Poisson distribution with mean L.

We first find an expression for the expected number of junctions present in a randomly selected chromosome. We explore the effects of population size, growth and type of growth, and develop two variance approximations. Under three specific assumptions, the variance is equal to the mean. Under these same assumptions, we show that the covariance is equal to the expected number of junctions shared between two chromosomes. This variance approximation is not good for very small populations, but demonstrates the aspects of the process which give rise to variation. The second approximation is obtained by relaxing two of the three assumptions, and performs well particularly in moderately large populations. In the second section, we find the expected number of junctions shared between two randomly selected chromosomes, and examine how this quantity is affected by population size and growth.

2.1 Number of Junctions per Morgan

2.1.1 Expectation

Suppose we randomly select a chromosome from the population at generation t. Let J_t be the number of junctions present on that chromosome. Let $\underline{N} = \{N_0, N_1, \ldots, N_t\}$, where N_j denotes the population size at time j, and let $\underline{n} = \{n_0, n_1, \ldots n_t\}$, where n_j denotes the number of junctions formed in meioses from generation j. Finally, let $\mathrm{I}_t(k,j) = 1$ if the kth junction formed in meioses from generation j is present on the chromosome selected at time t, and let $\mathrm{I}_t(k,j) = 0$ otherwise. Then

$$J_t \mid \underline{N}, \underline{n} = \sum_{j=0}^{t-1} \sum_{k=1}^{n_j} \mathrm{I}_t(k,j) \quad . \tag{2.1}$$

Taking the expectation over the indicator function,

$$\mathrm{E}_I \left[J_t \mid \underline{N}, \underline{n} \right] = \sum_{j=0}^{t-1} \sum_{k=1}^{n_j} \mathrm{E}[\mathrm{I}_t(k,j)] \quad . \tag{2.2}$$

Now $\mathrm{E}[\mathrm{I}_t(k,j)]$ =Pr(junction k from generation j is present on the selected chromosome). Let l denote the locus where junction k formed, and consider the population at generation $j+1$. We can think of locus l as having two alleles - one which is junction k, and one which is *not* junction k. The frequency of k in generation $j+1$ is exactly $\frac{1}{2N_{j+1}}$. Since we are considering a random mating population, each of the $2N_{j+1}$ genes at locus l in generation $j+1$ are equally likely to be the ancestor of locus l in the randomly selected chromosome. Therefore $\mathrm{E}[\mathrm{I}_t(k,j)]$ =Pr(junction k from generation j is present on the selected chromosome)= $\frac{1}{2N_{j+1}}$, and thus

$$\mathrm{E}_I \left[J_t \mid \underline{N}, \underline{n} \right] = \sum_{j=0}^{t-1} \frac{n_j}{2N_{j+1}} \quad . \tag{2.3}$$

If we now take the expectation over the junction formation process,

$$\mathrm{E}[J_t \mid \underline{N}] = \mathrm{E}_{\underline{n}} \left[\mathrm{E}_I [J_t \mid \underline{N}, \underline{n}] \right] = \sum_{j=0}^{t-1} \frac{\mathrm{E}[n_j]}{2N_{j+1}} \quad . \tag{2.4}$$

We show in Appendix B, Equation B.4 that

$$\mathrm{E}[n_j] = 2N_{j+1} h_j(\underline{N}) L \ , \tag{2.5}$$

where $h_j(\underline{N})$ is the probability of non-IBD at a particular locus in an individual in generation j, and L is the length of the chromosome in Morgans. Thus we obtain

$$\mathrm{E}[J_t \mid \underline{N}] = \sum_{j=0}^{t-1} h_j(\underline{N}) \cdot L \ . \tag{2.6}$$

For the random mating population considered here,

$$h_j(\underline{N}) = \prod_{i=0}^{j-1} \left(1 - \frac{1}{2N_i}\right) \ . \tag{2.7}$$

This result allows calculation of the number of junctions expected in a chromosome, as a function of population sizes, thereby allowing the exploration of the effects of different patterns of population growth.

Constant Population Size

We first consider a population of constant size N. In this case, $h_j(\underline{N}) = \left(1 - \frac{1}{2N}\right)^j$, and so

$$\begin{aligned} \mathrm{E}[J_t \mid \underline{N}] &= \sum_{j=0}^{t-1} \left(1 - \frac{1}{2N}\right)^j \cdot L \\ &= 2NL \cdot \left(1 - \left(1 - \frac{1}{2N}\right)^t\right) \\ &= 2NL \cdot (1 - h_t(\underline{N})) \ . \end{aligned} \tag{2.8}$$

Figure 2.1 shows the expected number of junctions per Morgan for small populations of constant size $N = 20$, 50 and 100. Note that as $t \to \infty$, $h_t(\underline{N}) \to 0$, and therefore $\mathrm{E}[J_t \mid \underline{N}] \to 2NL$. This result was previously outlined by Blossey [3]. Convergence is quicker for smaller populations, since $h_t(\underline{N})$ (the probability of non-IBD) goes to zero more quickly. Larger populations accumulate more junctions because regions of non-IBD, which are required for junction formation, are maintained for longer periods of time. In the limit as $t \to \infty$, as the populations become fixed, doubling the population size doubles the number of junctions expected, and therefore approximately halves the average length of intact ancestral segments.

It is interesting to note that $\mathrm{E}[J_t \mid \underline{N}]$ is linear in $h_t(\underline{N})$. Figure 2.2 shows the expected number of junctions per Morgan as a function of the inbreeding coefficient $(1 - h_t(\underline{N}))$.

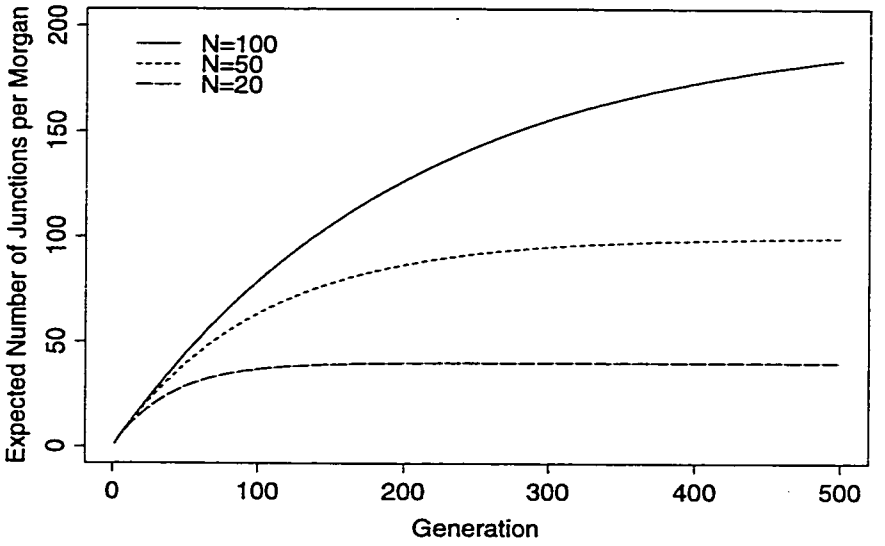


Figure 2.1: Expected number of junctions per Morgan, for a population of constant size N.

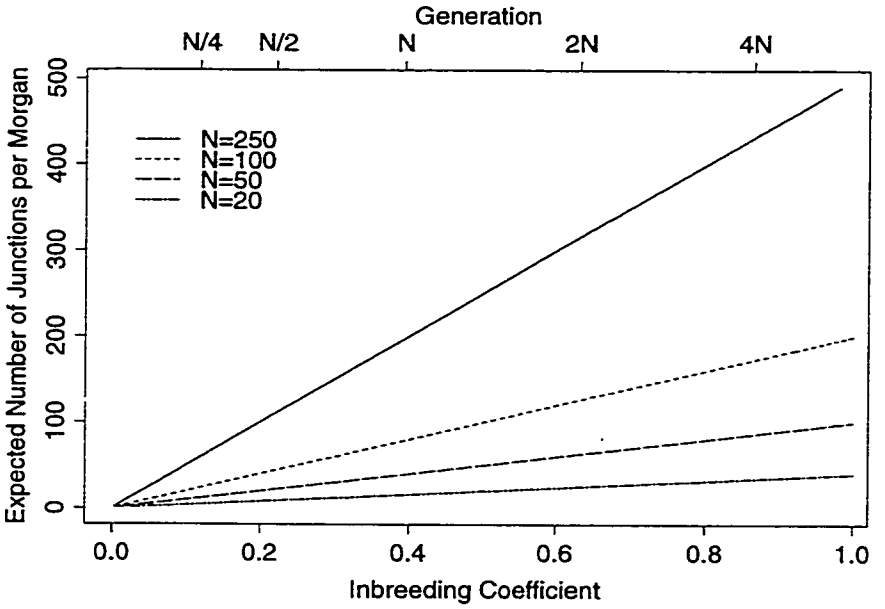


Figure 2.2: Expected number of junctions per Morgan, as a function of the inbreeding coefficient.

The top axis shows the generation number as a function of population size, using the approximation $h_t(\underline{N}) = (1 - \frac{1}{2N})^t \simeq \exp \frac{-t}{2N}$. This linear relationship between $\mathrm{E}[J_t \mid \underline{N}]$ and the inbreeding coefficient is only true for populations of constant size.

Growing Populations

We now consider growing populations, and compare the effects of exponential vs. linear growth on the the expected number of junctions present in a randomly selected chromosome from generation t. For the exponentially growing population,

$$N_t = N_0 \cdot g^t \ , \tag{2.9}$$

where g is the factor by which the population size is multiplied in each generation, and $g > 1$. Then,

$$h_j(\underline{N}) = \prod_{i=0}^{j-1} \left(1 - \frac{1}{2N_0 g^i}\right) \ . \tag{2.10}$$

Jacquard [17] shows that as $j \to \infty$, $h_j(\underline{N}) \to \alpha_g$, where $\alpha_g > 0$. This implies that $\mathrm{E}[J_t \mid \underline{N}]$ diverges. Junctions continue to accumulate indefinitely, in contrast to a population of constant size. The segments of ancestral chromosome therefore get smaller and smaller as the population continues to evolve.

In a linearly growing population,

$$N_t = N_0 + m \cdot t \ , \tag{2.11}$$

where m is the linear growth rate, and $m > 0$. Now

$$h_j(\underline{N}) = \prod_{i=0}^{j-1} \left(1 - \frac{1}{2N_0 + 2mi}\right) \ . \tag{2.12}$$

Jacquard [17] shows that as $j \to \infty$, $h_j(\underline{N}) \to 0$, although the convergence is extremely slow. It is therefore possible that $\mathrm{E}[J_t \mid \underline{N}]$ converges. We investigate the behaviour of $\mathrm{E}[J_t \mid \underline{N}]$ in the limit using Raabe's Test (see for example [19]). Let $R_j = j\left(1 - \frac{h_{j+1}(\underline{N})}{h_j(\underline{N})}\right)$. Then Raabe's Test states that:

(a) If $\liminf_{j\to\infty} R_j > 1$ then $\lim_{t\to\infty} \sum_{j=1}^{t} h_j(\underline{N})$ converges and

(b) If there exists some j^* such that $R_j \leq 1 \ \ \forall j \geq j^*$ then $\lim_{t\to\infty} \sum_{j=1}^{t} h_j(\underline{N})$ diverges.

For the linearly growing population,

$$R_j = j \cdot 1 - \left(1 - \frac{1}{2N_0 + 2mj}\right) = \frac{j}{2N_0 + 2mj} \ . \tag{2.13}$$

If $m \geq \frac{1}{2}$, then $R_j \leq 1 \ \ \forall j \geq 1$, so $\mathrm{E}[J_t \mid \underline{N}]$ diverges by part (b) above. Now,

$$\liminf_{j\to\infty} R_j = \liminf_{j\to\infty} \frac{j}{2N_0 + 2mj} = \frac{1}{2m} \ , \tag{2.14}$$

and so for $m < \frac{1}{2}$, $\mathrm{E}[J_t \mid \underline{N}]$ converges by part (a) above. Thus for linear growth rates greater than or equal to $\frac{1}{2}$, the expected number of junctions diverges, and in this sense, the linearly growing population is similar to the exponentially growing population. For very small linear growth rates ($m < \frac{1}{2}$), the expected number of junctions does converge, as it does in a population of constant size.

Figure 2.3 shows some examples of the effect of small linear growth rates on $\mathrm{E}[J_t \mid \underline{N}]$ per Morgan for four types of population, each with a starting population size (N_0) of 20, and with different growth rates. For the population of constant size ($m = 0$), the expected number of junctions per Morgan converges to 40. The populations with $m = 0.1$ and $m = 0.2$, corresponding approximately to the addition of one individual to the population every 10 and 5 generations respectively, are also converging, and are not substantially different from the constant sized population. The expected number of junctions in the population with $m = 1$ continues to increase, and is dramatically different from the populations with smaller growth rates. These examples are not meant to represent the situation in human populations, but rather to illustrate how important growth rate can be. After 500 generations, a chromosome sampled from a population adding one extra individual per generation ($m = 1$) is expected to have over four times as many junctions as a chromosome sampled from a population of constant size. This implies that intact ancestral segments are only one quarter as long in the growing population, even though the growth is very slow.

Example populations with different types of growth

We next consider a somewhat more realistic population model, and compare the effects of linear vs. exponential growth at finite times. We consider populations with founding size

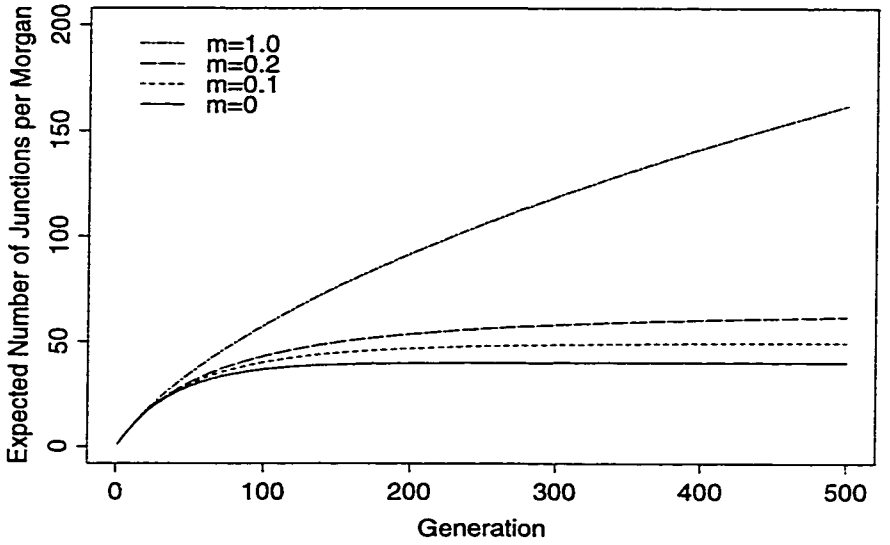


Figure 2.3: Expected number of junctions per Morgan, for populations in which $N_0 = 20$, and $m = 0$, 0.1, 0.2 or 1.0.

$N_0 = 20$, which have grown to a size of $N_{100} = 2020$ after one hundred generations. This time-depth reflects that of the population of Finland, which was founded approximately 2000 years ago [25]. We compare the expected number of junctions per Morgan in populations who achieved this growth (i) linearly, with $m = 20$ and (ii) exponentially, with $g = 1.047233$. Figure 2.4 shows the population sizes over time corresponding to the two types of growth. Exponential growth of a population can be hard to maintain due to scarcity of resources after a number of generations have passed. The linear growth model is also imperfect, in that it implies decreasing fertility per-individual over time. Neither of these models are necessarily reflective of any human population, but it is illustrative to examine the effects of the two different growth patterns.

Figure 2.5 shows the expected number of junctions per Morgan in a randomly selected chromosome for populations growing linearly and exponentially, for founding population sizes (a) $N_0 = 20$, (b) $N_0 = 100$ and (c) $N_0 = 500$. For a given founding population size, the populations have the same size at generations 0 and 100, but their trajectories

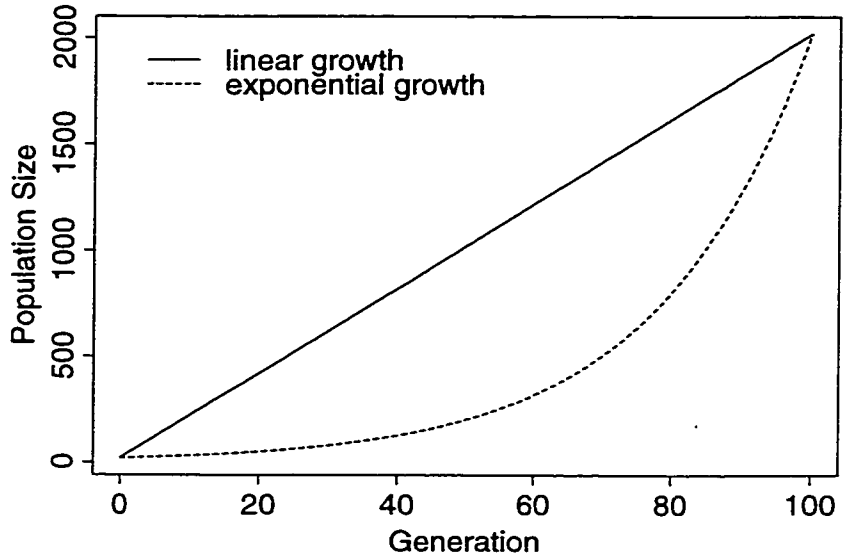


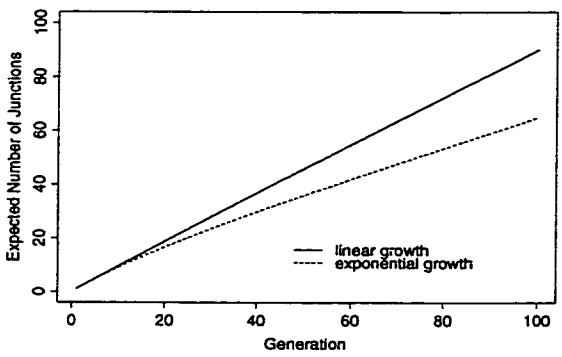
Figure 2.4: Population sizes for linear and exponential growth, $N_0 = 20$ and $N_{100} = 2020$.

between the two times are different (see Figure 2.4). Fewer junctions are expected in the exponentially growing population because the population remains quite small for a long period of time. During this time, the expected proportion of the chromosome which is non-IBD is greatly reduced, and therefore the number of junctions formed in later meioses is greatly reduced.

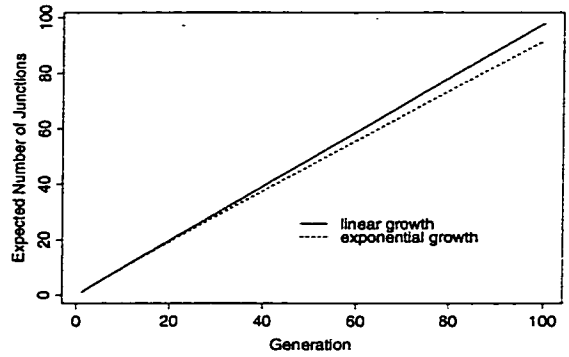
Table 2.1 shows the expected number of junctions per Morgan at time 100, for linear and exponential growth and three different founding population sizes. For the smallest

Table 2.1: Expected number of junctions per Morgan in a chromosome taken from generation 100, for exponential and linear growth.

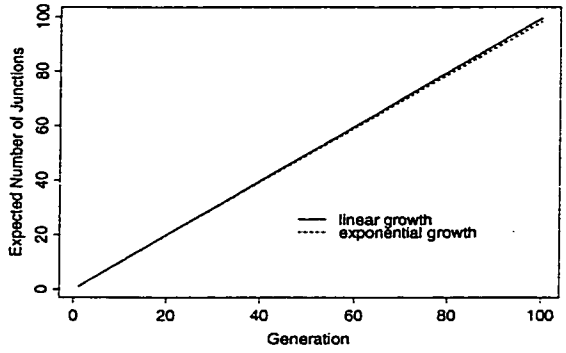
	(N_0)		
Growth Type	20	100	500
linear growth	90.0	97.9	99.6
exponential growth	64.9	91.7	98.3



(a) $N_0 = 20$



(b) $N_0 = 100$



(c) $N_0 = 500$

Figure 2.5: Expected number of junctions per Morgan as function of generation, for populations growing exponentially and linearly. Founder population sizes are (a) 20, (b) 100, and (c) 500.

populations, one expects approximately 90 junctions in a chromosome sampled from the linearly growing population, and approximately 65 junctions in a chromosome sampled from the exponentially growing population. In this case, there are 40% more junctions in the linearly growing population, and therefore the intact ancestral segments are expected to be much smaller on average. This effect of growth type is quite strong in the smallest populations, but barely noticeable in the largest populations.

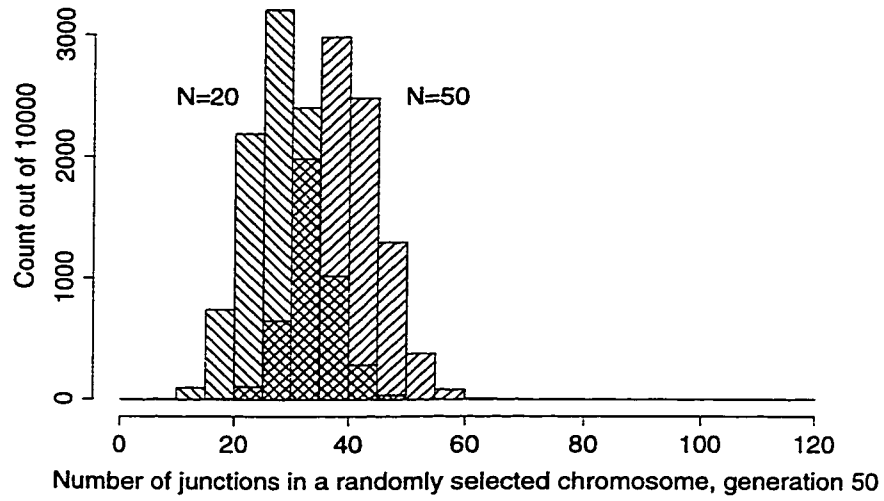
2.1.2 Variance

We demonstrated in the previous section that population size, growth rate, and type of growth can have a strong effect on the expected number of junctions in a randomly chosen chromosome. In natural populations we observe only one "replicate" of a population, and if the process of junction formation and transmission is very variable, the number of junctions existing in a chromosome sampled from that replicate may be quite different from the number expected. For this reason, it is important that we also consider the variance of $J_t \mid \underline{N}$.

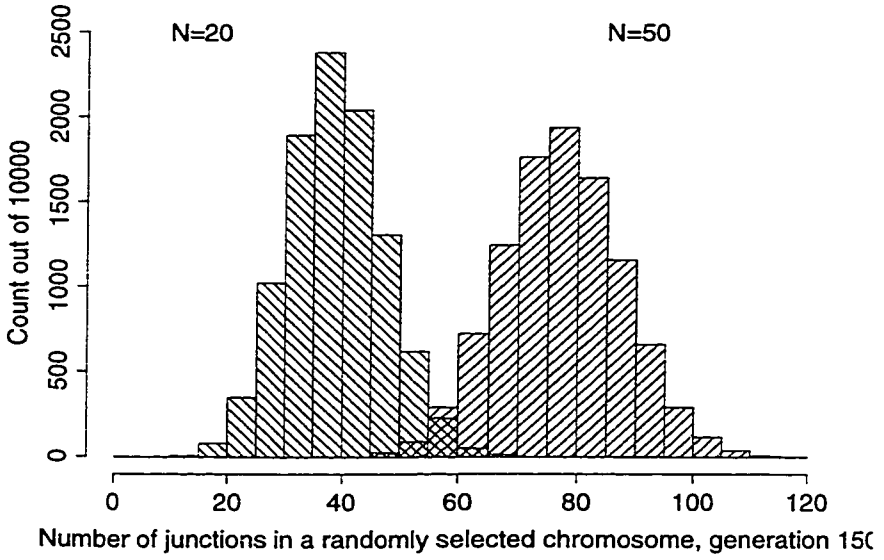
Figure 2.6 shows the empirical distribution, based on 10,000 simulations, of $J_t \mid N$, for populations of constant size $N = 20$ and $N = 50$, at generation numbers 50 and 150. When the populations are younger, there is a large amount of overlap between the two distributions, even though the means are quite different ($\mathrm{E}[J_{50} \mid N = 20] = 28.27$, $\mathrm{E}[J_{50} \mid N = 50] = 39.50$), reflecting that $Var[J_{50} \mid N]$ is large. For the older populations, the variance remains large, but there is enough separation between the means that there is very little overlap of the distributions. This example demonstrates the need for at least an approximation to the variance of $J_t \mid \underline{N}$.

We first consider an approximation based on some simplifying assumptions. The approximation does not work well, but it is informative to consider the sources of variation in $J_t \mid \underline{N}$. Suppose that

1. The number of junctions (n_i) formed in meioses from generation i has a Poisson distribution with mean $2N_{i+1}h_i(\underline{N})L$. This would be true if all of the individuals in generation i had the same proportion $h_i(\underline{N})$ of their genome non-IBD. In reality,



(a) t=50



(b) t=150

Figure 2.6: Empirical distributions of $J_{50} \mid N$ and $J_{150} \mid N$, for $N = 20$ and $N = 50$.

this proportion varies across members of generation i, and is equal to $h_i(\underline{N})$ only in expectation. This extra variability leads to extra-Poisson variation in the distribution of n_i.

2. The number of junctions (n_i) formed in meioses from generation i is independent of the number of junctions (n_j) formed in meioses from generation j, for all $i \neq j$. This is not generally true. For example, knowing that n_i is very small relative to the number of meioses implies that the population is likely close to fixation, and therefore subsequent n_j ($j > i$) must also be small.

3. The presence of any one junction in the sampled chromosome from generation t is independent of the presence of any other junction in that chromosome. That is *Pr(junction k formed in a meiosis from generation i exists in the chromosome sampled at generation t | junction l formed in a meiosis from generation j exists in the chromosome sampled at generation t)= Pr(junction k formed in a meiosis from generation i exists in the chromosome sampled at generation t), for any k, l, i and j, where* $k \neq l$ *if* $i = j$.

Let $J_t(i)$ denote the number of junctions formed in generation i which exist in the randomly sampled chromosome from generation t. Then assumption 1, together with the fact that the probability that a junction formed in a meiosis from generation i exists in the chromosome sampled at generation t equals $\frac{1}{2N_{i+1}}$, implies that $J_t(i) \mid \underline{N} \sim Poisson(h_i(\underline{N})L)$, $0 \leq i \leq t-1$. Furthermore, assumptions 2 and 3 imply that

$$J_t(i) \mid \underline{N} \text{ is independent of } J_t(j) \mid \underline{N}\ , \text{ for } i \neq j\ . \tag{2.15}$$

Therefore

$$J_t \mid \underline{N} = \sum_{j=0}^{t-1} \{J_t(j) \mid \underline{N}\} \sim Poisson\left(\sum_{j=0}^{t-1} h_j(\underline{N})L\right)\ . \tag{2.16}$$

For the Poisson distribution, the variance is equal to the mean, and can therefore be calculated using Equation 2.6.

Consideration of the simulations presented in Figure 2.6 immediately demonstrates the inadequacy of this approximation to the variance of $J_t \mid \underline{N}$. The following table shows the

ratio of the estimated variance ($\hat{\sigma}^2$, based on 10,000 simulations) to the theoretical mean(μ) as a function of N and t:

$\hat{\sigma}^2$ / μ	$t = 20$	t=50	$t = 125$	t=150
N=20	1.08	1.28	1.70	1.76
N=50	0.99	1.06	1.29	1.34

The Poisson variance is an underestimate of the true variance, especially for older generations, and the smaller population. Comparing the populations at times where t/N is equal to one ($N = 20$, $t = 20$ and $N = 50$, $t = 50$) we see that the mean underestimates the variance by approximately the same amount: 8 or 9 percent. Similarly, comparing populations where $t/N = 1.5$ ($N = 20$, $t = 50$ and $N = 50$, $t = 125$) the mean underestimates the variance by 28 or 29 percent. This suggests that for a constant sized population, t/N approximately determines the adequacy of the Poisson approximation to the variance. For values of t/N larger than 1, the Poisson approximation underestimates the true variance. The importance of the quantity t/N is not surprising, since for a population of constant size, $h_t(\underline{N}) = (1 - \frac{1}{2N})^t \simeq \exp \frac{-t}{2N}$. Larger values of t/N correspond to increasing amounts of IBD in the population, and in these situations, assumptions 1 through 3 may be further from the truth.

Covariance

Before we develop an alternative variance approximation, we first find a simple expression for the covariance under the assumptions discussed above. Let $J_t^{(1)}$ denote the number of junctions in the first sampled chromosome, and let $J_t^{(2)}$ denote the number of junctions in a second chromosome, sampled without replacement of the first. Then

$$\begin{aligned} J_t^{(1)} J_t^{(2)} \mid \underline{N}, \underline{n} &= \sum_{j=0}^{t-1} \sum_{k=1}^{n_j} \mathrm{I}_t^{(1)}(k,j) \cdot \sum_{j'=0}^{t-1} \sum_{l=1}^{n_{j'}} \mathrm{I}_t^{(2)}(l,j') \\ &= \sum_{j=0}^{t-1} \sum_{\substack{j'=0, \\ j' \neq j}}^{t-1} \sum_{k=1}^{n_j} \sum_{l=1}^{n_{j'}} \mathrm{I}_t^{(1)}(k,j) \cdot \mathrm{I}_t^{(2)}(l,j') \\ &+ \sum_{j=0}^{t-1} \sum_{k=1}^{n_j} \sum_{\substack{l=1 \\ l \neq k}}^{n_j} \mathrm{I}_t^{(1)}(k,j) \cdot \mathrm{I}_t^{(2)}(l,j) + \end{aligned}$$

$$+ \sum_{j=0}^{t-1} \sum_{k=1}^{n_j} \mathrm{I}_t^{(1)}(k,j) \cdot \mathrm{I}_t^{(2)}(k,j) \ . \tag{2.17}$$

Note that the first term is a sum over junctions formed in different generations, the second term is a sum over different junctions formed in the same generation, and the third term is over all junctions. We will take the same approach as before, taking the expectation first with respect to the indicator variables, and second with respect to the $\underline{n}$. We require another assumption, which is the analogue of assumption 3 above, namely that the presence of any junction in chromosome one is independent of the presence or absence of any other junction in chromosome two. This assumption gives us independence of the indicator variables in terms one and two. In term three, the product of the indicators is simply an indicator of whether or not the junction in question is present in both chromosomes one and two. We will denote this indicator by $I_t^B(k,j)$. Since $\mathrm{E}_I[\mathrm{I}_t^{(1)}(k,j)] = \mathrm{E}_I[\mathrm{I}_t^{(2)}(k,j)] = \frac{1}{2N_{j+1}}$ for all k and all j, we have

$$\begin{aligned}
\mathrm{E}_I[J_t^{(1)} J_t^{(2)} \mid \underline{N}, \underline{n}] \ &= \ \sum_{j=0}^{t-1} \sum_{\substack{j'=0, \\ j' \neq j}}^{t-1} \sum_{k=1}^{n_j} \sum_{l=1}^{n_{j'}} \frac{1}{2N_{j+1}} \cdot \frac{1}{2N_{j'+1}} \\
&+ \ \sum_{j=0}^{t-1} \sum_{k=1}^{n_j} \sum_{\substack{l=1 \\ l \neq k}}^{n_j} \frac{1}{2N_{j+1}} \cdot \frac{1}{2N_{j+1}} + \\
&+ \ \sum_{j=0}^{t-1} \sum_{k=1}^{n_j} \mathrm{E}_I[I_t^B(k,j)] \ .
\end{aligned} \tag{2.18}$$

Note that $\mathrm{E}_I[I_t^B(k,j)]$ depends only on t and j. We let $p_t^B(j)$ denote $\mathrm{E}_I[I_t^B(k,j)]$, and then taking expectations with respect to the $\underline{n}$, we obtain

$$\mathrm{E}[J_t^{(1)} J_t^{(2)} \mid \underline{N}] = \sum_{j=0}^{t-1} \sum_{\substack{j'=0, \\ j' \neq j}}^{t-1} \frac{\mathrm{E}[n_j n_{j'}]}{2N_{j+1} 2N_{j'+1}} + \sum_{j=0}^{t-1} \frac{\mathrm{E}[n_j(n_j-1)]}{2N_{j+1} 2N_{j+1}} + \sum_{j=0}^{t-1} \mathrm{E}[n_j] \cdot p_t^B(j). \tag{2.19}$$

To find the covariance, we must subtract the product of the expected values:

$$\mathrm{E}[J_t^{(1)} \mid \underline{N}] \cdot \mathrm{E}[J_t^{(2)} \mid \underline{N}] = \left(\sum_{j=0}^{t-1} \frac{\mathrm{E}[n_j]}{2N_{j+1}} \right)^2 = \sum_{j=0}^{t-1} \sum_{\substack{j'=0 \\ j' \neq j}}^{t-1} \frac{\mathrm{E}[n_j]\mathrm{E}[n_{j'}]}{2N_{j+1} 2N_{j'+1}} + \sum_{j=0}^{t-1} \frac{\mathrm{E}[n_j]^2}{(2N_{j+1})^2} \ . \tag{2.20}$$

Note that under assumption 2 (independence of n_j and $n_{j'}$), the first term of Equation 2.20 is equal to the first term of Equation 2.19. Under assumption 1 (n_j has a Poisson distribution),

$\mathrm{E}[n_j(n_j-1)] = \mathrm{E}[n_j^2] - \mathrm{E}[n_j] = \mathrm{E}[n_j]^2$, and so the second term of Equation 2.20 is equal to the second term of Equation 2.19. Therefore, under these assumptions,

$$Cov(J_t^{(1)}, J_t^{(2)}) = \sum_{j=0}^{t-1} \mathrm{E}[n_j] \cdot p_t^B(j) \ . \tag{2.21}$$

Note that this is just the expected number of junctions shared between the two chromosomes. We will discuss the calculation of this quantity in Section 2.2.

We now develop a second variance approximation, which does not require assumptions 1 and 2. Consider the calculation of $\mathrm{E}[J_t^2 \mid \underline{N}]$. Now,

$$\begin{aligned} J_t^2 \mid \underline{N}, \underline{n} &= \left(\sum_{j=0}^{t-1} \sum_{k=1}^{n_j} \mathrm{I}_t(k,j) \right)^2 \\ &= \sum_{j=0}^{t-1} \left\{ \sum_{k=1}^{n_j} \mathrm{I}_t(k,j) \right\}^2 + \sum_{i=0}^{t-1} \sum_{\substack{j=0, \\ j \neq i}}^{t-1} \left\{ \sum_{k=1}^{n_i} \mathrm{I}_t(k,i) \right\} \left\{ \sum_{l=1}^{n_j} \mathrm{I}_t(k,l) \right\} \\ &= \sum_{j=0}^{t-1} \sum_{k=1}^{n_j} \mathrm{I}_t(k,j) + \sum_{j=0}^{t-1} \sum_{k=1}^{n_j} \sum_{\substack{l=1, \\ l \neq k}}^{n_j} \mathrm{I}_t(k,j)\mathrm{I}_t(l,j) \\ &\quad + \sum_{i=0}^{t-1} \sum_{\substack{j=0, \\ j \neq i}}^{t-1} \sum_{k=1}^{n_i} \sum_{l=1}^{n_j} \mathrm{I}_t(k,i)\mathrm{I}_t(l,j) \ . \end{aligned} \tag{2.22}$$

The first term in $J_t^2 \mid \underline{N}, \underline{n}$ is a sum over all junctions. The second term is a sum over pairs of distinct junctions formed in the same generation, and the third term is a sum over pairs of junctions formed in different generations. We use the same approach of taking successive expectations as we did in calculating $\mathrm{E}[J_t \mid \underline{N}]$. Then

$$\begin{aligned} \mathrm{E}[J_t^2 \mid \underline{N}] &= \mathrm{E}_{\underline{n}}[\mathrm{E}_I[J_t^2 \mid \underline{N}, \underline{n}]] \\ &= \mathrm{E}_{\underline{n}} \left[\sum_{j=0}^{t-1} \sum_{k=1}^{n_j} \mathrm{E}[\mathrm{I}_t(k,j)] \right. \\ &\quad + \sum_{j=0}^{t-1} \sum_{k=1}^{n_j} \sum_{\substack{l=1, \\ l \neq k}}^{n_j} \mathrm{E}[\mathrm{I}_t(k,j)\mathrm{I}_t(l,j)] \\ &\quad \left. + \sum_{i=0}^{t-1} \sum_{\substack{j=0, \\ j \neq i}}^{t-1} \sum_{k=1}^{n_i} \sum_{l=1}^{n_j} \mathrm{E}[\mathrm{I}_t(k,i)\mathrm{I}_t(l,j)] \right] \ . \end{aligned} \tag{2.23}$$

We showed previously that $\mathrm{E}[\mathrm{I}_t(k,j)] = \frac{1}{2N_{j+1}}$. Using assumption 3, we have:

$$\mathrm{E}[\mathrm{I}_t(k,j)\mathrm{I}_t(l,j)] \simeq \frac{1}{2N_{j+1}} \cdot \frac{1}{2N_{j+1}}, \quad \mathrm{E}[\mathrm{I}_t(k,i)\mathrm{I}_t(l,j)] \simeq \frac{1}{2N_{i+1}} \cdot \frac{1}{2N_{j+1}}. \tag{2.24}$$

Then

$$\begin{aligned}
\mathrm{E}[J_t^2 \mid \underline{N}] &= \mathrm{E}_{\underline{n}}[\mathrm{E}_I[J_t^2 \mid \underline{N}, \underline{n}]] \\
&\simeq \mathrm{E}_{\underline{n}}\left[\sum_{j=0}^{t-1}\sum_{k=1}^{n_j}\frac{1}{2N_{j+1}}\right] + \mathrm{E}_{\underline{n}}\left[\sum_{j=0}^{t-1}\sum_{k=1}^{n_j}\sum_{\substack{l=1,\\ l\neq k}}^{n_j}\frac{1}{4N_{j+1}^2}\right] \\
&\quad + \mathrm{E}_{\underline{n}}\left[\sum_{i=0}^{t-1}\sum_{\substack{j=0,\\ j\neq i}}^{t-1}\sum_{k=1}^{n_i}\sum_{l=1}^{n_j}\frac{1}{4N_{i+1}N_{j+1}}\right]
\end{aligned} \tag{2.25}$$

and

$$\mathrm{E}[J_t^2 \mid \underline{N}] \simeq \sum_{j=0}^{t-1}\frac{\mathrm{E}[n_j]}{2N_{j+1}} + \sum_{j=0}^{t-1}\frac{\mathrm{E}[n_j(n_j-1)]}{4N_{j+1}^2} + \sum_{i=0}^{t-1}\sum_{\substack{j=0,\\ j\neq i}}^{t-1}\frac{\mathrm{E}[n_i n_j]}{4N_{i+1}N_{j+1}}. \tag{2.26}$$

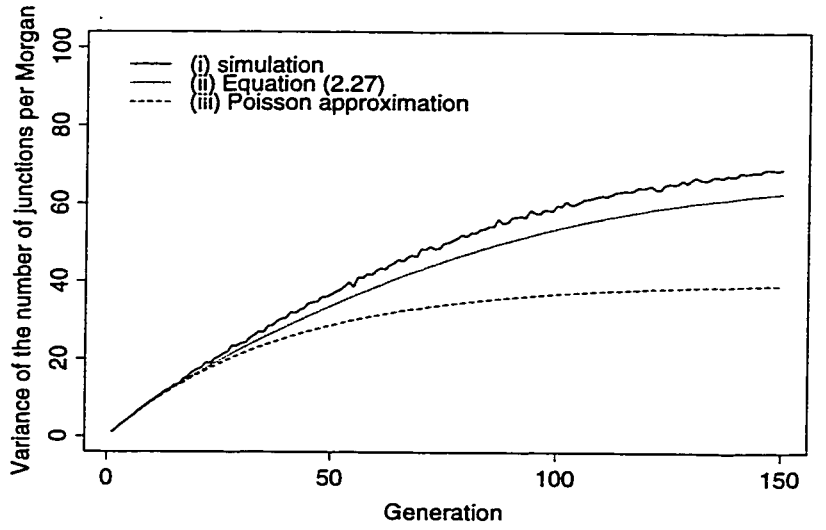
Expressions for $\mathrm{E}[n_j^2]$ and $\mathrm{E}[n_i n_j]$ are developed in Appendix B, Equations B.7 and B.9. We can now approximate the variance of J_t by

$$\begin{aligned}
Var[J_t \mid \underline{N}] &= \mathrm{E}[J_t^2 \mid \underline{N}] - \mathrm{E}[J_t \mid \underline{N}]^2 \\
&\simeq \sum_{j=0}^{t-1}\frac{\mathrm{E}[n_j]}{2N_{j+1}} + \sum_{j=0}^{t-1}\frac{\mathrm{E}[n_j(n_j-1)]}{4N_{j+1}^2} + \\
&\quad \sum_{i=0}^{t-1}\sum_{\substack{j=0,\\ j\neq i}}^{t-1}\frac{\mathrm{E}[n_i n_j]}{4N_{i+1}N_{j+1}} - \left(\sum_{j=0}^{t-1}\frac{\mathrm{E}[n_j]}{2N_{j+1}}\right)^2.
\end{aligned} \tag{2.27}$$

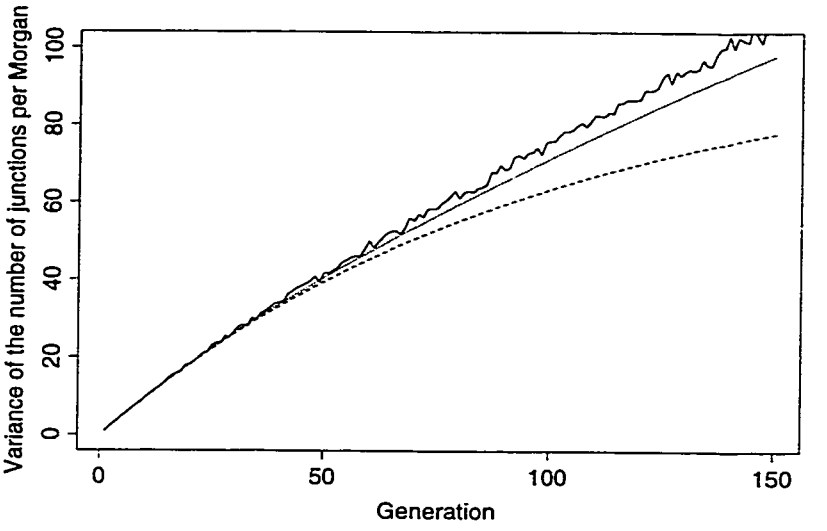
The expectations in this expression depend on the chromosome length, and the population sizes over time, through the single locus non-IBD probabilities ($h_t(\underline{N})$), and the two locus non-IBD probabilities ($\Theta_t(\underline{N},\theta)$, $\Gamma_t(\underline{N},\theta)$, $\Delta_t(\underline{N},\theta)$), which are described in Appendix A.

Constant Population Size

Figure 2.7 shows the variance of the number of junctions per Morgan in populations of constant size either $N = 20$ or $N = 50$. The variance is estimated by simulation (10,000



(a) $N = 20$



(b) $N = 50$

Figure 2.7: Variance of the number of junctions in a randomly selected chromosome from a population of constant size $N = 20$ or $N = 50$; estimated by (i) simulation, (ii) Equation 2.27, and (iii) the Poisson approximation.

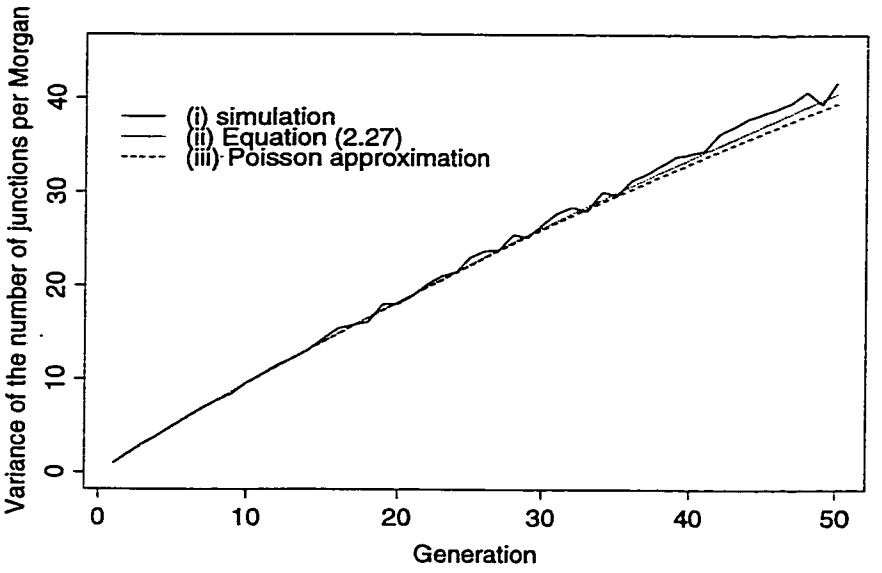


Figure 2.8: Variance of the number of junctions in early generations from a population of constant size $N = 50$; estimated by (i) simulation, (ii) Equation 2.27, and (iii) the Poisson approximation.

iterations), Equation 2.27 and the Poisson approximation. For both populations, Equation 2.27 is much better than the Poisson based variance approximation, particularly for later generations. It is interesting to note that in both examples, the Poisson approximation begins to fail at approximately the Nth generation, which is where the non-IBD proportion has been reduced to about 60% (see Figure 2.2). For generations earlier than this, the two variance approximations are almost indistinguishable, and they are very close to the simulated variance (see Figure 2.8). This suggests that for young populations, or older, larger populations, the Poisson variance approximation may be adequate. The Poisson approximation to the variance has an advantage over Equation 2.27, because it is so much easier to calculate.

Example populations with different types of growth

We now return to the example considered in Section 2.1.1 of populations growing either exponentially or linearly to 100 times their founding population size over 100 generations.

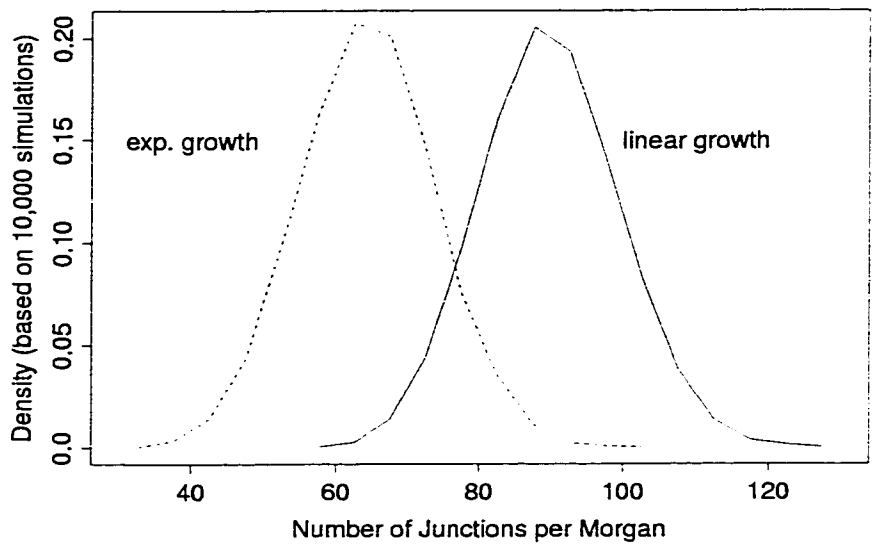


Figure 2.9: Distributions (estimated from 10,000 simulations) of the number of junctions in a randomly selected chromosome from generation 100, for populations growing either linearly or exponential with $N_0 = 20$.

Figure 2.9 shows the distributions estimated from 10,000 simulations of the number of junctions in a chromosome of length one Morgan, randomly selected from the population at generation 100, for populations growing either linearly or exponentially, with $N_0 = 20$. Both distributions have a very large variance and there is considerable overlap between them. This shows that while the expected number of junctions per Morgan may be quite different between two populations, the number of junctions actually observed in two populations with different growth patterns could be quite similar by chance. This demonstrates the importance of being able to quantify the variance of the number of junctions in a chromosome.

Table 2.2 quantifies these observations. The table shows the variance of the number of junctions in a chromosome of length one Morgan from generation 100, estimated by (i) simulation, (ii) the Poisson approximation, and (iii) Equation 2.27. Simulation based estimates are available only for populations with $N_0 = 20$, since simulation of the larger populations is too computationally demanding. For the exponentially growing population with $N_0 = 20$,

Table 2.2: Variance of the number of junctions in a chromosome randomly selected from generation 100, based on 10,000 simulations, the Poisson approximation, or Equation 2.27.

	$N_0 = 20$		$N_0 = 100$		$N_0 = 500$	
	exponential	linear	exponential	linear	exponential	linear
simulation	86.94	92.04	-	-	-	-
Poisson	64.93	90.04	91.74	97.93	98.29	99.58
Equation 2.27	80.50	90.68	92.06	97.93	98.29	99.58

the Poisson approximation badly underestimates the variance. The approximation based on Equation 2.27 is much better, but still an underestimate. Both approximations are much better for the linearly growing population with $N_0 = 20$. For the larger populations, the Poisson approximation and Equation 2.27 give very similar results, and we hypothesize that the variance is well approximated in these populations.

2.2 *Number of junctions shared per Morgan*

In this section we develop an expression for the expected value of the number of junctions shared between two randomly selected chromosomes from generation t, as a function of the population sizes over time. We apply this result to both populations of constant size, and growing populations.

2.2.1 Expected number shared between two chromosomes

We again consider a random mating population, and study S_t, the number of junctions shared between two chromosomes randomly selected without replacement from generation t. Taking the approach of summing over the generation of formation,

$$S_t \mid \underline{N}, \underline{n} = \sum_{j=0}^{t-1} \sum_{k=1}^{n_j} \mathrm{I}_t^B(k,j) \quad . \tag{2.28}$$

In this equation, $\mathrm{I}_t^B(k,j)$ indicates that both of the selected chromosomes carry the kth junction formed in meioses from generation j. Taking the expectation with respect to the

indicator function, we require $\mathrm{E}[\mathrm{I}_t^B(k,j)]$, which is simply the probability that both of two chromosomes selected from generation t carry the kth junction formed in meioses from generation j. If we let Z_t be the number of copies of that junction existing in the selected chromosomes (i.e. Z_t = 0,1 or 2), and Y_i be the number of copies of that junction that are present in generation i, then the required probability is $\Pr(Z_t = 2 \mid Y_{j+1} = 1)$. Taking successive expectations, as in Equation 2.4, we obtain

$$\mathrm{E}[S_t \mid \underline{N}] = \sum_{j=0}^{t-2} \mathrm{E}[n_j] \cdot Pr(Z_t = 2 \mid Y_{j+1} = 1) \ . \tag{2.29}$$

The summation only goes as high as $t-2$, since $Pr(Z_t = 2 \mid Y_t = 1) = 0$.

Calculation of $Pr(Z_t = 2 \mid Y_{j+1} = 1)$

For this random mating population, $\{Y_{j+1}, Y_{j+2}, \ldots\}$ is a Markov chain in which $Y_{j+1} = 1$,

$$Y_i \mid Y_{i-1} \sim \mathrm{Bin}\left(2\mathrm{N_i}, \frac{\mathrm{Y_{i-1}}}{2\mathrm{N_{i-1}}}\right) \quad \mathrm{i} \geq \mathrm{j}+2, \tag{2.30}$$

and conditional on Y_{i-1}, Y_i is independent of Y_l, for all $l < i-1$. Similarly

$$Z_t \mid Y_{t-1} \sim \mathrm{Bin}\left(2, \frac{\mathrm{Y_{t-1}}}{2\mathrm{N_{t-1}}}\right) \ \mathrm{t} \geq \mathrm{j}+2, \tag{2.31}$$

and conditional on Y_{t-1}, Z_t is independent of Y_l for all $l < i-1$. Therefore

$$\begin{aligned} Pr\left(Z_t = 2 \mid Y_{j+1} = 1\right) &= \mathrm{E}\left[Pr\left(Z_t = 2 \mid Y_{t-1}, Y_{j+1} = 1\right) \mid Y_{j+1} = 1\right] \\ &= \mathrm{E}\left[Pr\left(Z_t = 2 \mid Y_{t-1}\right) \mid Y_{j+1} = 1\right] \\ &= \mathrm{E}\left[\left(\frac{Y_{t-1}}{2N_{t-1}}\right)^2 \mid Y_{j+1} = 1\right] \\ &= \frac{1}{(2N_{t-1})^2}\mathrm{E}\left[Y_{t-1}^2 \mid Y_{j+1} = 1\right] \ t \geq j+2. \end{aligned} \tag{2.32}$$

In order to find an expression for $\mathrm{E}\left[Y_{t-1}^2 \mid Y_{j+1} = 1\right]$, we will need the following results. The binomial distribution of $Y_i \mid Y_{i-1}$ implies that

$$\mathrm{E}\left[Y_i \mid Y_{i-1}\right] = \frac{N_i}{N_{i-1}} Y_{i-1} \quad i \geq j+2, \tag{2.33}$$

and

$$\mathrm{E}\left[Y_i^2 \mid Y_{i-1}\right] = Y_{i-1}^2 \left(\frac{N_i}{N_{i-1}}\right)^2 \left(1 - \frac{1}{2N_i}\right) + Y_{i-1}\frac{N_i}{N_{i-1}} \quad i \geq j+2. \tag{2.34}$$

Also,

$$\begin{aligned} \mathrm{E}[Y_i \mid Y_{j+1}=1] &= \mathrm{E}[\mathrm{E}[Y_i \mid Y_{i-1}] \mid Y_{j+1}=1] \\ &= \mathrm{E}\left[\frac{N_i}{N_{i-1}} Y_{i-1} \mid Y_{j+1}=1\right] \\ &= \frac{N_i}{N_{i-1}} \mathrm{E}[Y_{i-1} \mid Y_{j+1}=1] \\ &= \frac{N_i}{N_{j+1}} \quad i \geq j+2. \end{aligned} \tag{2.35}$$

We will now show by induction that

$$\mathrm{E}\left[Y_{t-1}^2 \mid Y_{j+1}=1\right] = \sum_{i=1}^{t-j-2} \left(\frac{N_{t-1}}{N_{j+i}}\right)^2 \frac{N_{j+i}}{N_{j+1}} \prod_{r=i}^{t-j-2} \left(1 - \frac{1}{2N_{j+r+1}}\right) + \frac{N_{t-1}}{N_{j+1}} \quad t \geq j+3\ . \tag{2.36}$$

When $t = j+3$,

$$\mathrm{E}\left[Y_{j+2}^2 \mid Y_{j+1}=1\right] = \left(\frac{N_{j+2}}{N_{j+1}}\right)^2 \left(1 - \frac{1}{2N_{j+2}}\right) + \frac{N_{j+2}}{N_{j+1}}\ , \tag{2.37}$$

by Equation 2.34, so Equation 2.36 holds for $t = j+3$. Now suppose that Equation 2.36 holds for $t = n$, and consider $\mathrm{E}[Y_{t-1}^2 | Y_{j+1}=1]$ for $t = n+1$.

$$\begin{aligned} \mathrm{E}\left[Y_n^2 | Y_{j+1}=1\right] &= \mathrm{E}\left[\mathrm{E}\left[Y_n^2 \mid Y_{n-1}\right] \mid Y_{j+1}=1\right] \\ &= \mathrm{E}\left[\left(Y_{n-1}^2 \left(\frac{N_n}{N_{n-1}}\right)^2 \left(1 - \frac{1}{2N_n}\right) + Y_{n-1} \frac{N_n}{N_{n-1}}\right) \mid Y_{j+1}=1\right] \\ &= \left(\frac{N_n}{N_{n-1}}\right)^2 \left(1 - \frac{1}{2N_n}\right) \mathrm{E}\left[Y_{n-1}^2 \mid Y_{j+1}\right] + \frac{N_n}{N_{n-1}} \mathrm{E}[Y_{n-1} \mid Y_{j+1}=1] \\ &= \left(\frac{N_n}{N_{n-1}}\right)^2 \left(1 - \frac{1}{2N_n}\right) \cdot \sum_{i=1}^{n-j-2} \left(\frac{N_{n-1}}{N_{j+i}}\right)^2 \frac{N_{j+i}}{N_{j+1}} \prod_{r=i}^{n-j-2} \left(1 - \frac{1}{2N_{j+r+1}}\right) \\ &+ \left(\frac{N_n}{N_{n-1}}\right)^2 \left(1 - \frac{1}{2N_n}\right) \cdot \frac{N_{n-1}}{N_{j+1}} + \frac{N_n}{N_{n-1}} \cdot \frac{N_{n-1}}{N_{j+1}}\ , \end{aligned} \tag{2.38}$$

using Equations 2.36 with $t = n$ (by supposition) and 2.35 in the last step. Simplifying,

$$\begin{aligned} \mathrm{E}\left[Y_n^2 | Y_{j+1}=1\right] &= \sum_{i=1}^{n-j-2} \left(\frac{N_n}{N_{j+i}}\right)^2 \frac{N_{j+i}}{N_{j+1}} \prod_{r=i}^{n-j-1} \left(1 - \frac{1}{2N_{j+r+1}}\right) \\ &+ \left(\frac{N_n}{N_{n-1}}\right)^2 \cdot \frac{N_{n-1}}{N_{j+1}} \cdot \left(1 - \frac{1}{2N_n}\right) + \frac{N_n}{N_{j+1}} \\ &= \sum_{i=1}^{n-j-1} \left(\frac{N_n}{N_{j+i}}\right)^2 \frac{N_{j+i}}{N_{j+1}} \prod_{r=i}^{n-j-1} \left(1 - \frac{1}{2N_{j+r+1}}\right) + \frac{N_n}{N_{j+1}}\ . \end{aligned} \tag{2.39}$$

Equation 2.39 is equivalent to Equation 2.36 with $t = n+1$, so we have shown by induction that Equation 2.36 holds for all $t \geq j+3$. Therefore

$$Pr(Z_t \mid Y_{j+1} = 1) = \begin{cases} \frac{1}{(2N_{t-1})^2}\left[\sum_{i=1}^{t-j-2}\left(\frac{N_{t-1}}{N_{j+i}}\right)^2 \frac{N_{j+i}}{N_{j+1}} \prod_{r=i}^{t-j-2}\left(1-\frac{1}{2N_{j+r+1}}\right) + \frac{N_{t-1}}{N_{j+1}}\right] & j \leq t-3 \\ \frac{1}{(2N_{t-1})^2} & j = t-2, \end{cases} \tag{2.40}$$

since $\mathrm{E}[Y_{t-1}^2 \mid Y_{t-1} = 1] = 1$. Substitution of these quantities and the expression for $\mathrm{E}[n_j]$ developed in Appendix B, into Equation 2.29 allows the calculation of $\mathrm{E}[S_t \mid \underline{N}]$. Thus we can calculate the expected number of junctions shared between two randomly selected chromosomes, as a function of the population sizes over time.

Constant Population Size

For a population of constant size, the equations can be simplified. For example, for a population of constant size N,

$$\begin{aligned} \mathrm{E}[Y_{t-1}^2 \mid Y_{j+1} = 1] &= \sum_{i=1}^{t-j-2} \prod_{r=i}^{t-j-2}\left(1-\frac{1}{2N}\right) + 1 \\ &= \sum_{i=1}^{t-j-2}\left(1-\frac{1}{2N}\right)^{t-j-i-1} + 1 \\ &= \sum_{i=0}^{t-j-2}\left(1-\frac{1}{2N}\right)^{i} \\ &= 2N\left[1-\left(1-\frac{1}{2N}\right)^{t-j-1}\right] \quad j \leq t-3. \end{aligned} \tag{2.41}$$

Then

$$Pr(Z_t \mid Y_{j+1} = 1) = \begin{cases} \frac{1}{(2N)^2} 2N\left[1-\left(1-\frac{1}{2N}\right)^{t-j-1}\right] & j \leq t-3 \\ \frac{1}{(2N)^2} & j = t-2, \end{cases} \tag{2.42}$$

since $\mathrm{E}[Y_{t-1}^2 \mid Y_{t-1} = 1] = 1$. Recall that $\mathrm{E}[n_j] = 2Nh_j(\underline{N})L$, and so by substitution of the above equations into Equation 2.29, we obtain

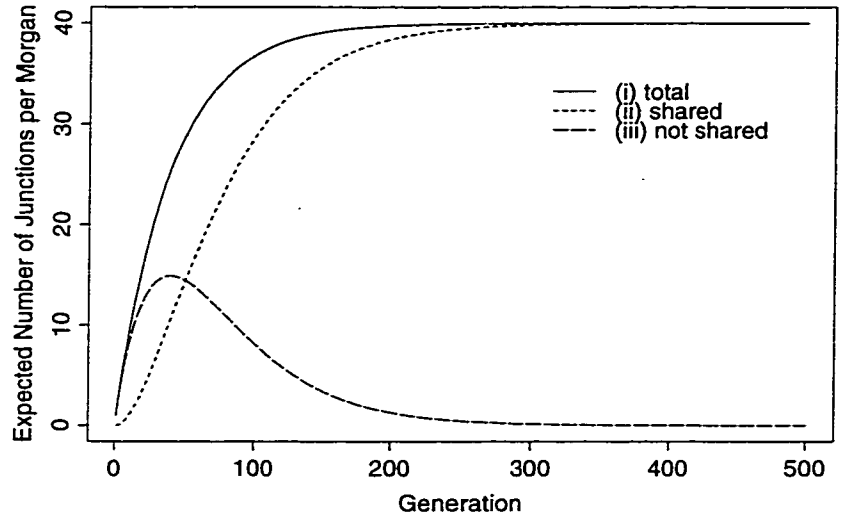
$$\begin{aligned} \mathrm{E}[S_t \mid \underline{N}] &= \sum_{j=0}^{t-2} \mathrm{E}[n_j] \cdot Pr(Z_t = 2 \mid Y_{j+1} = 1) \\ &= \sum_{j=0}^{t-3} 2Nh_j(\underline{N})L \cdot \frac{1}{(2N)^2} 2N\left[1-\left(1-\frac{1}{2N}\right)^{t-j-1}\right] + 2Nh_{t-2}(\underline{N})L \cdot \frac{1}{(2N)^2} \end{aligned}$$

$$
\begin{aligned}
&= L\left\{\sum_{j=0}^{t-3} h_j(\underline{N}) \cdot \left[1-\left(1-\frac{1}{2N}\right)^{t-j-1}\right] + h_{t-2}(\underline{N}) \cdot \frac{1}{2N}\right\} \\
&= L\left\{\sum_{j=0}^{t-2} h_j(\underline{N}) \cdot \left[1-\left(1-\frac{1}{2N}\right)^{t-j-1}\right]\right\} \\
&= L\left\{\sum_{j=0}^{t-2} \left(1-\frac{1}{2N}\right)^{j} \cdot \left[1-\left(1-\frac{1}{2N}\right)^{t-j-1}\right]\right\} \\
&= L\left\{\sum_{j=0}^{t-2} \left(1-\frac{1}{2N}\right)^{j} - (t-1)\left(1-\frac{1}{2N}\right)^{t-1}\right\} \\
&= L\left\{2N\left(1-\left(1-\frac{1}{2N}\right)^{t-1}\right) - (t-1)\left(1-\frac{1}{2N}\right)^{t-1}\right\} \\
&= L\left\{2N\left(1-h_{t-1}(\underline{N})\right) - (t-1)h_{t-1}(\underline{N})\right\} \\
&= 2NL\left\{1-h_{t-1}(\underline{N}) + \frac{1}{2N}h_{t-1}(\underline{N}) - \frac{t}{2N}h_{t-1}(\underline{N})\right\} \\
&= 2NL\left\{1-h_{t-1}(\underline{N})\left(1-\frac{1}{2N}\right) - \frac{t}{2N}h_{t-1}(\underline{N})\right\} \\
&= 2NL\left\{1-h_t(\underline{N}) - t\left[h_{t-1}(\underline{N}) - h_t(\underline{N})\right]\right\}, \qquad (2.43)
\end{aligned}
$$

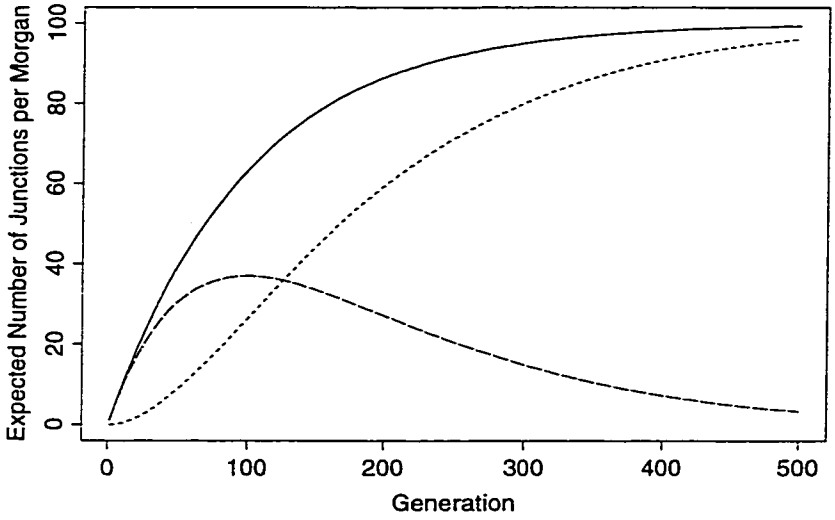
since $h_t(\underline{N}) = \left(1-\frac{1}{2N}\right)^t = \left(1-\frac{1}{2N}\right)\cdot h_{t-1}(\underline{N})$. Note that as $t \to \infty$, $t\left[h_{t-1}(\underline{N}) - h_t(\underline{N})\right] \to 0$, and therefore $\mathrm{E}[J_t \mid N]$ and $\mathrm{E}[S_t \mid N]$ approach the same limit. To understand this result, it is easiest to think of randomly selecting a chromosome, comparing it to its homologue in the same individual, and interpreting $\mathrm{E}[S_t \mid N]$ as the expected number of junctions IBD between the two chromosomes. This interpretation is valid, since in a randomly mating population, the two chromosomes within an individual are equivalent to two randomly selected chromosomes. As the population goes to fixation and $\mathrm{E}[J_t \mid N]$ goes to a finite limit, all junctions in the two chromosomes must be shared, since a junction which is not shared indicates a region of non-IBD. Therefore it follows that $\mathrm{E}[J_t \mid N]$ and $\mathrm{E}[S_t \mid N]$ should approach the same limit. The difference between J_t and S_t can be interpreted as the number of junctions present in the selected chromosome which are *not* shared IBD with its homologue, and we denote this quantity U_t. Note that

$$\mathrm{E}[U_t \mid N] = \mathrm{E}[J_t \mid N] - \mathrm{E}[S_t \mid N] = 2NL \cdot t\left[h_{t-1}(\underline{N}) - h_t(\underline{N})\right] \qquad (2.44)$$

Figure 2.10 shows $\mathrm{E}[J_t \mid N]$, $\mathrm{E}[S_t \mid N]$ and $\mathrm{E}[U_t \mid N]$, for a population of constant size $N = 20$ or $N = 50$, evolving over 500 generations. In earlier generations, the population is



(a) $N = 20$



(b) $N = 50$

Figure 2.10: Expected number of junctions in a randomly selected chromosome from a population of constant size $N = 20$ or $N = 50$; (i) total, (ii) shared IBD with its homologue, and (iii) not shared IBD with its homologue.

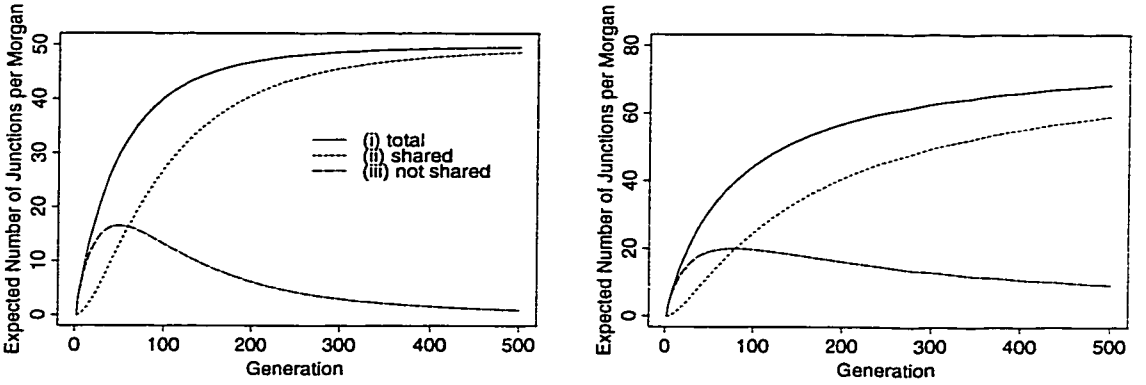
far from fixation - many junctions have been formed, but few have reached a high enough frequency to be shared IBD between homologues. During this time period, most of the junctions present in the chromosome are not shared. As the population continues to evolve, the IBD proportion increases - fewer junctions are formed, and more are fixed or lost due to drift. Eventually, most of the junctions present in the chromosome are shared IBD with its homologue, and in the limit, all are. The total number of junctions becomes dominated by the shared junctions more quickly in the smaller population. This is because junctions formed in the early generations reach higher frequencies more quickly than they do in the larger population, where their initial frequency is much lower. This is an example of the more pronounced effect of genetic drift in the smaller population.

Growing Populations

We again consider populations which are growing linearly or exponentially. Figure 2.11 shows $\mathrm{E}[J_t \mid \underline{N}]$, $\mathrm{E}[S_t \mid \underline{N}]$ and $\mathrm{E}[U_t \mid \underline{N}]$, for a population with $N_0 = 20$ and linear growth rates $m = 0.1$, $m = 0.25$ or $m = 1.0$. For the two cases where $m < 0.5$, we showed previously that $\mathrm{E}[J_t \mid \underline{N}]$ converges to a finite number. Since populations growing in this manner are known to go to fixation as $t \to \infty$ [17], this implies that $\mathrm{E}[U_t|\underline{N}] \to 0$, and $\mathrm{E}[S_t \mid \underline{N}]$ must approach the same limit as $\mathrm{E}[J_t \mid \underline{N}]$. This is clearly the case in panels (a) and (b). Panel (c) shows the case where $m = 1.0$. We saw previously that $\mathrm{E}[J_t \mid \underline{N}]$ diverges in this situation, and since the population must go to fixation [17], this implies that $\mathrm{E}[S_t \mid \underline{N}]$ must also diverge, while $\mathrm{E}[U_t \mid \underline{N}]$ may diverge more slowly, or converge to a finite number of junctions.

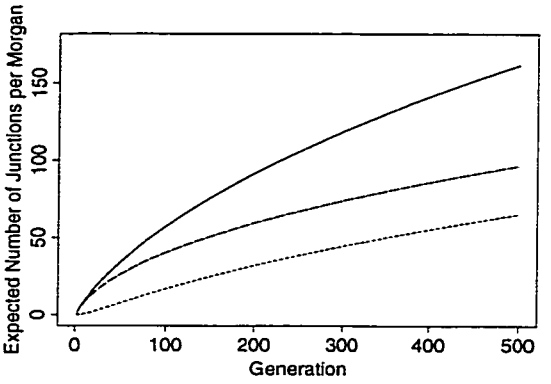
Example populations with different types of growth

Recall the example previously considered, of a population growing either linearly or exponentially to 100 times its size at founding over 100 generations, as shown in Figure 2.4. Figures 2.12, 2.13 and 2.14 show the expected total number of junctions in a randomly selected chromosome, decomposed into those shared IBD with its homologous chromosome in the same individual, and those which are not shared IBD for the populations with founding



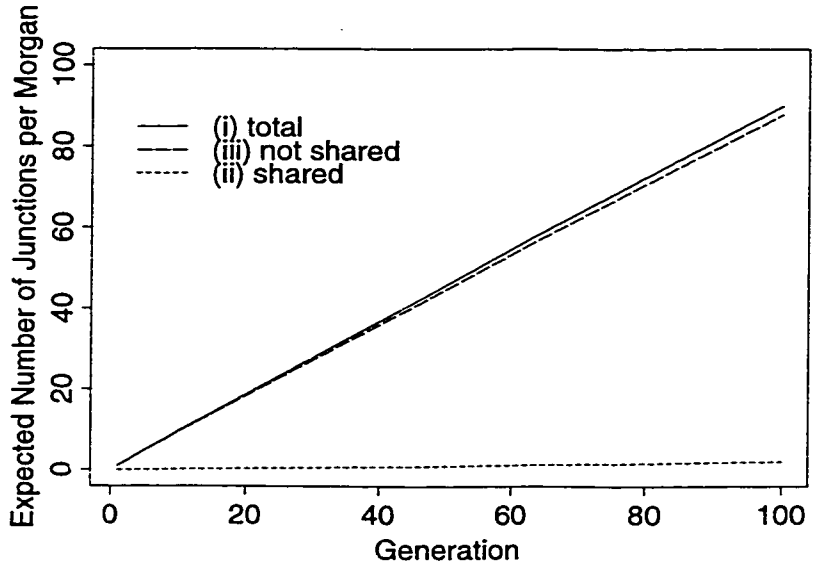
(a) m=0.10

(b) m=0.25

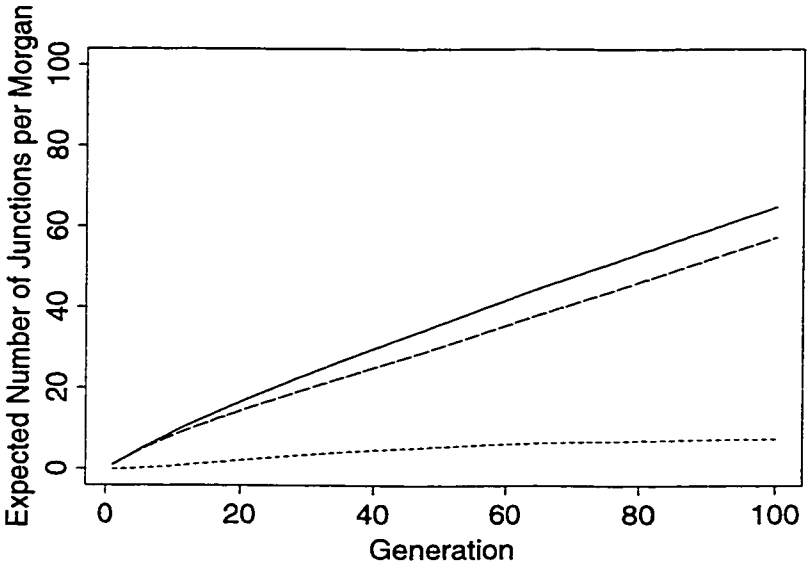


(c) m=1.00

Figure 2.11: Expected number of junctions in a randomly selected chromosome from a linearly growing population with $N_0 = 20$ and $m = 0.10$, $m = 0.25$ or $m = 1.00$; (i) total, (ii) shared IBD with its homologue, and (iii) not shared IBD with its homologue.

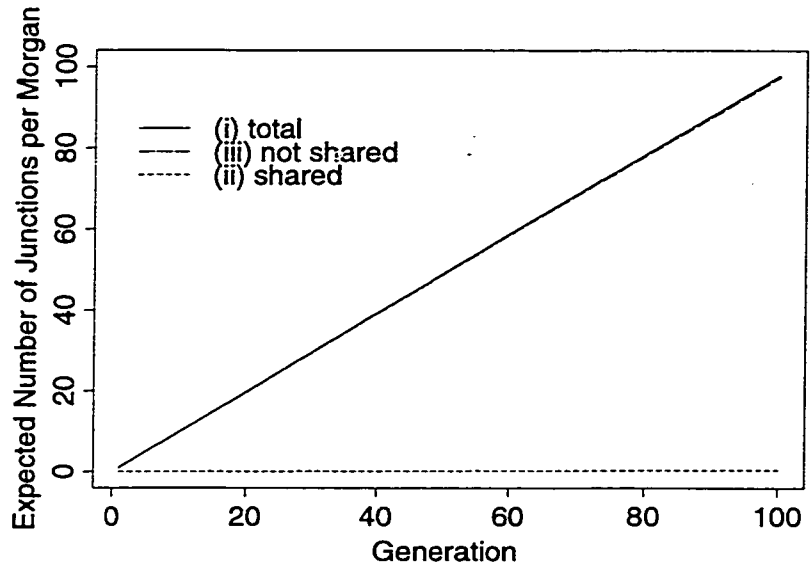


(a) Linear Growth

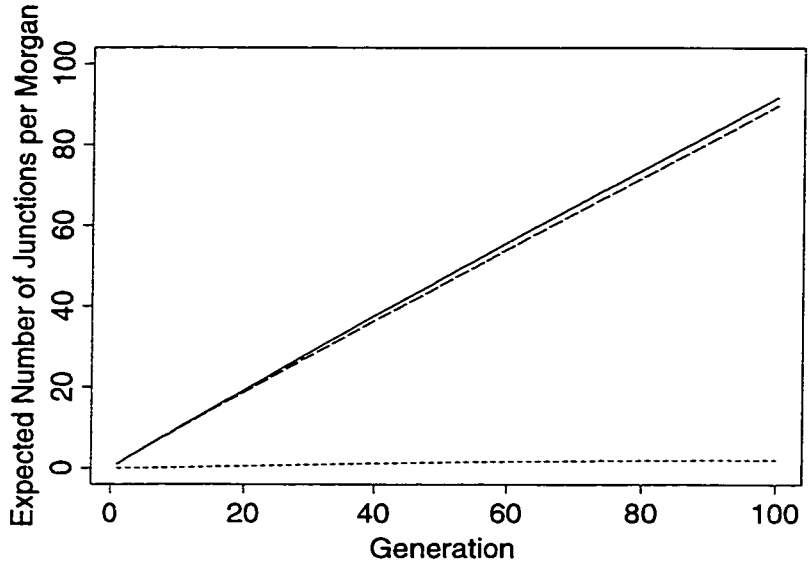


(b) Exponential Growth

Figure 2.12: Expected number of junctions in a randomly selected chromosome from a population growing either linearly or exponentially from $N_0 = 20$; (i) total, (ii) shared IBD with its homologue, and (iii) not shared IBD with its homologue.

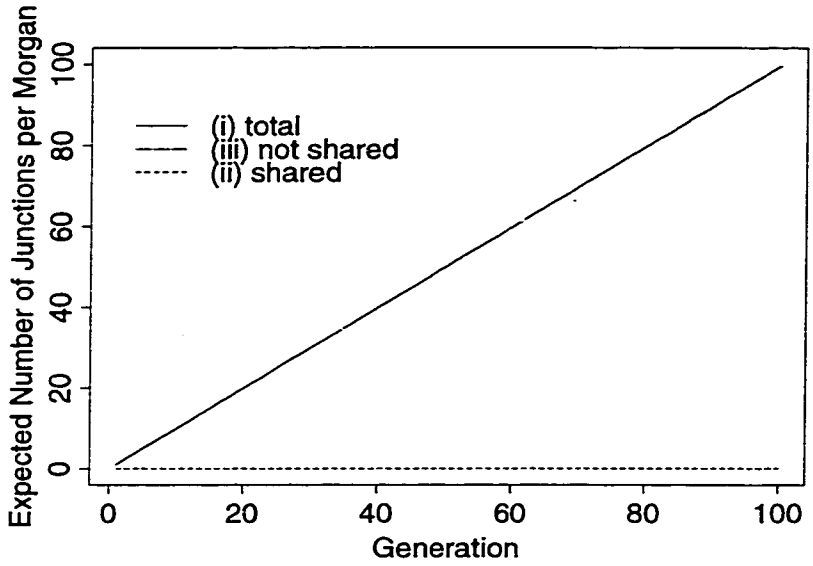


(a) Linear Growth

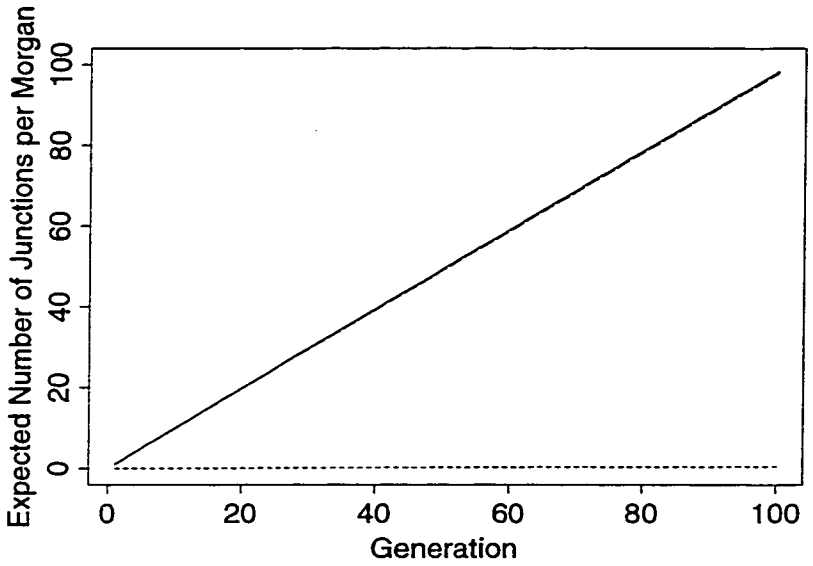


(b) Exponential Growth

Figure 2.13: Expected number of junctions in a randomly selected chromosome from a population growing either linearly or exponentially from $N_0 = 100$; (i) total, (ii) shared IBD with its homologue, and (iii) not shared IBD with its homologue.



(a) Linear Growth



(b) Exponential Growth

Figure 2.14: Expected number of junctions in a randomly selected chromosome from a population growing either linearly or exponentially from $N_0 = 500$; (i) total, (ii) shared IBD with its homologue, and (iii) not shared IBD with its homologue.

sizes $N_0 = 20$, $N_0 = 100$ and $N_0 = 500$ respectively. In the exponentially growing populations, a greater fraction of the expected total number of junctions are expected to be shared IBD between the two chromosomes. This reflects the fact that the exponentially growing populations maintain smaller sizes for a longer time. During this period, the frequencies of junctions that were formed in the earliest generations drift such that some of them are in relatively high frequencies, and they are therefore more likely to be shared by two homologous chromosomes. In contrast, the effects of drift are much less pronounced in the linearly growing populations, which do not remain small for long. Junctions formed in these populations have a much lower initial frequency, and it therefore takes much longer for their frequencies to drift high enough for them to be shared by two homologous chromosomes. This example demonstrates that different patterns of growth affect not only the total number of junctions in a chromosome, but also how these junctions are distributed, in terms of how many are shared IBD with the homologous chromosome in the same individual, and how many are not.

2.3 Discussion

In this chapter, we developed an approach to calculating the expected number of junctions in a chromosome sampled from the present generation of an isolated population. We developed some theoretical results for the limiting number of junctions in populations of constant size, and for populations growing linearly or exponentially. We applied our approach to some example populations, and find that for very small populations, different growth patterns can result in dramatic differences in the expected numbers of junctions in a chromosome. However, different growth patterns have a smaller effect in larger populations. We further demonstrated that the number of junctions in a chromosome has a high variance, and therefore considered two variance approximations. Under certain assumptions, the variance is equal to the mean, and the covariance of the numbers of junctions in two distinct chromosomes is equal to the expected number of junctions shared between the two. We showed how to calculate the expected number of junctions shared, and demonstrated that it can be affected by type of growth but only when populations are very small to begin

with.

Chapter 3

THE EFFECTS OF RANDOM SUBDIVISION AND NON-RANDOM MATING ON JUNCTION NUMBER AND SHARING

In this chapter, we study some departures from the previous model of a single non-subdivided random mating population. We study the effects of random subdivision on junction number and junction sharing in a random mating population, assuming that subpopulations arise as a result of random subdivision of the main population at a particular generation, and that there is no migration between subpopulations after the division has occurred. We also present some theoretical results on regular mating systems.

3.1 Population subdivision

In this section, we demonstrate how the results of Chapter 2 can be applied to subdivided populations. Specifically, we consider the mean and variance of the number of junctions in a randomly selected chromosome from a particular generation, and the expected number of junctions shared between two chromosomes either from within the same subpopulation or from two different subpopulations. These results are applied to some example populations similar to the ones considered in Chapter 2.

3.1.1 Expected number of junctions per Morgan

To calculate $\mathrm{E}[J_t \mid \underline{N}]$ for a chromosome randomly sampled from a particular subpopulation, we use Equation 2.6;

$$\mathrm{E}[J_t \mid \underline{N}] = \sum_{j=0}^{t-1} h_j(\underline{N}) \cdot L \ . \tag{3.1}$$

Recall that $\underline{N}$ is the vector $N_0, N_1, \ldots, N_t$ of population sizes. For generations prior to the subdivision of the population, N_i is the size of the entire population, and for generations

after the subdivision, N_i is the size of the subpopulation from which the chromosome is being sampled.

Constant Size Population

We now consider a population of constant total size N, which divides into m subpopulations of equal size at generation t_s. That is, at time t_s, the population is randomly divided into m groups of equal size, which reproduce independently of one another in subsequent generations. Then

$$\begin{aligned} h_i(\underline{N}) &= \left(1-\frac{1}{2N}\right)^i \quad 0 \le i \le t_s \text{ and} \\ h_{t_s+i}(\underline{N}) &= h_{t_s}(\underline{N}) \cdot \left(1-\frac{m}{2N}\right)^i \quad i = 1,2,\ldots \end{aligned} \tag{3.2}$$

We will use J_t^s to denote the number of junctions in a chromosome sampled from a subpopulation. Then

$$\begin{aligned} \mathrm{E}[J_t^s \mid N] &= \sum_{j=0}^{t-1} h_j(\underline{N})L \\ &= \sum_{j=0}^{t-1} \left(1-\frac{1}{2N}\right)^j L \\ &= 2NL \cdot \left[1-\left(1-\frac{1}{2N}\right)^t\right] \\ &= 2NL \cdot [1-h_t(\underline{N})], \quad t \le t_s+1. \end{aligned} \tag{3.3}$$

For $t > t_s + 1$,

$$\begin{aligned} \mathrm{E}[J_t^s \mid N] &= \sum_{j=0}^{t-1} h_j(\underline{N})L \\ &= \sum_{j=0}^{t_s} h_i(\underline{N})L + \sum_{i=1}^{t-t_s-1} h_{t_s+i}(\underline{N})L \\ &= 2NL\left(1-\left(1-\frac{1}{2N}\right)^{t_s+1}\right) + h_{t_s}(\underline{N}) \sum_{i=1}^{t-t_s-1} \left(1-\frac{m}{2N}\right)^i L \\ &= 2NL\left(1-\left(1-\frac{1}{2N}\right)^{t_s+1}\right) + \left(1-\frac{1}{2N}\right)^{t_s}\left(2\frac{N}{m}\left[1-\left(1-\frac{m}{2N}\right)^{t-t_s}\right]-1\right)L \\ &= 2NL\left(1-\left(1-\frac{1}{2N}\right)^{t_s+1} + \left(1-\frac{1}{2N}\right)^{t_s}\left(\frac{1}{m}\left[1-\left(1-\frac{m}{2N}\right)^{t-t_s}\right]-\frac{1}{2N}\right)\right) \end{aligned}$$

$$
\begin{aligned}
&=2NL\left(1-\left(1-\frac{1}{2N}\right)^{t_s}\left[1-\frac{1}{2N}-\frac{1}{m}\left[1-\left(1-\frac{m}{2N}\right)^{t-t_s}\right]+\frac{1}{2N}\right]\right)\\
&=2NL\left(1-\left(1-\frac{1}{2N}\right)^{t_s}\left[1-\frac{1}{m}\left[1-\left(1-\frac{m}{2N}\right)^{t-t_s}\right]\right]\right)\\
&=2NL\left(1-\left(1-\frac{1}{2N}\right)^{t_s}\left(1-\frac{1}{m}\right)-\left(1-\frac{1}{2N}\right)^{t_s}\frac{1}{m}\left(1-\frac{m}{2N}\right)^{t-t_s}\right)\\
&=2NL\left(1-h_{t_s}(\underline{N})\left[1-\frac{1}{m}+\frac{1}{m}\left(1-\frac{m}{2N}\right)^{t-t_x}\right]\right)\ . \qquad (3.4)
\end{aligned}
$$

Note that as $t \to \infty$,

$$
\mathrm{E}[J_t^s|\underline{N}] \to 2NL\cdot\left\{1-h_{t_s}(\underline{N})\left(1-\frac{1}{m}\right)\right\}\ . \qquad (3.5)
$$

Thus for a given number of subdivisions, earlier subdivision leads to fewer junctions being formed. This is because IBD accumulates faster in the smaller populations that exist after the subdivision. Similarly, for a given time of subdivision, more subdivision leads to smaller subpopulations and hence fewer junctions.

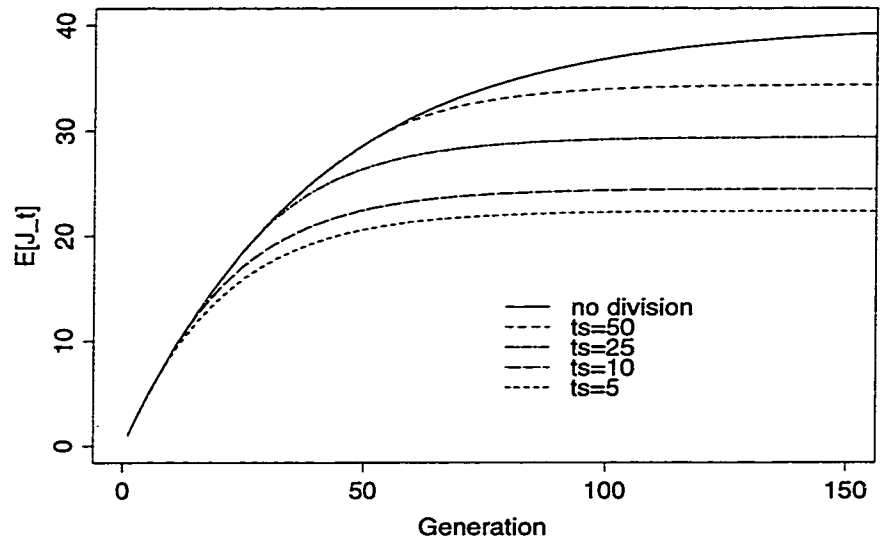
Figure 3.1 shows the effects of subdivision into two subpopulations at different times, for (a) N=20 and (b) N=50. The effects of subdivision are less pronounced in the larger of the two populations, and the differences are large only if the population has been reproducing for more than 50 generations. These results suggest that in order for subdivision to seriously impact the number of junctions in chromosomes in human populations, the populations must be initially very small, or extremely subdivided.

Example of growing populations with repeated subdivision

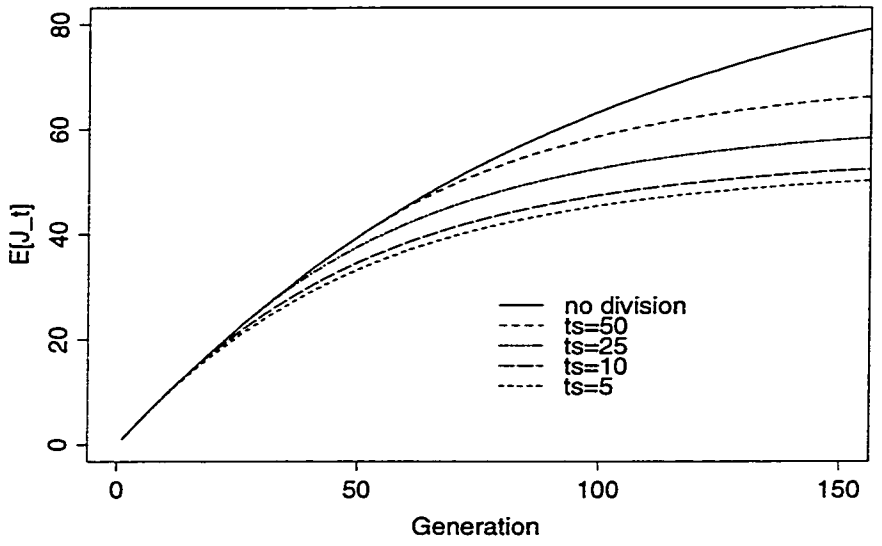
We now consider the effects of subdivision in an example which is more relevant to human populations. We again consider a population which grows exponentially to 100 times its initial size over a period of 100 generations, with initial population sizes (N_0) of 20, 100 and 500 individuals. We calculate $\mathrm{E}[J_{100}^s]$ for the following scenarios:

- an exponentially growing population without subdivision. A 100 fold increase over 100 generations corresponds to an exponential growth rate of 4.7% per generation.

- an exponentially growing population in which splits occur whenever a population size of $8N_0$ is reached.



(a) N=20



(b) N=50

Figure 3.1: Expected number of junctions in a population of size (a) N=20 or (b) N=50 divided into 2 subpopulations at varying times.

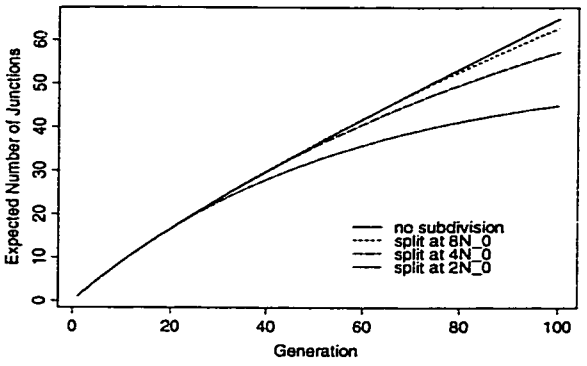
- an exponentially growing population in which splits occur whenever a population size of $4N_0$ is reached.

- an exponentially growing population in which splits occur whenever a population size of $2N_0$ is reached.

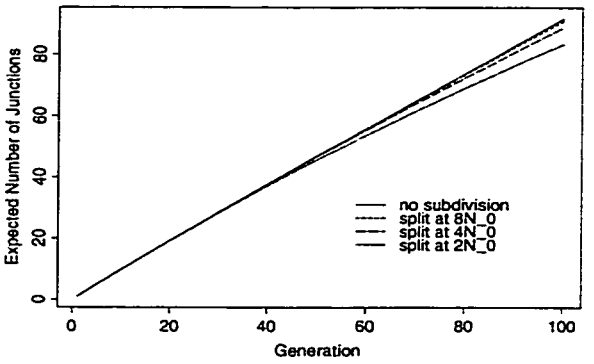
For a given value of N_0, all scenarios have the same total population size (summing over subpopulations) at every generation. The difference between the scenarios is in the amount of subdivision within the population. The most subdivided population (where splits occur when a size of $2N_0$ is reached) is made up of 64 subpopulations at time 100. The populations in which splits occur at $4N_0$ and $8N_0$ are made up of 32 and 16 subpopulations at time 100, respectively.

Figure 3.2 shows the expected number of junctions per Morgan for populations growing exponentially at a rate of 4.7% per generation. The different panels correspond to founding population sizes of (a) 20 (b) 100 and (c) 500, and different lines on the same plot correspond to different degrees of subdivision. For a given founding population size, the expected number of junctions in a particular generation increases as the amount of internal subdivision decreases. The differences are quite pronounced for the smallest population, and barely noticeable for the largest.

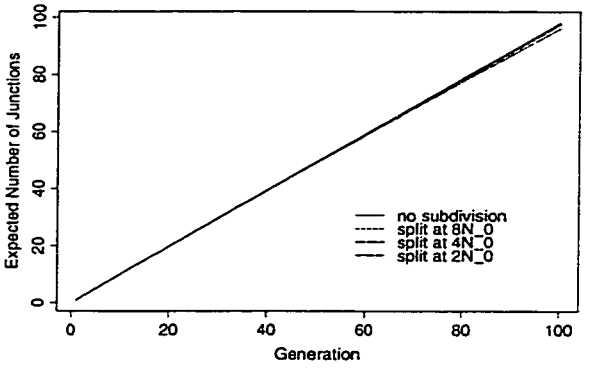
Table 3.1 shows the expected number of junctions per Morgan in a chromosome taken from generation 100, for the four different subdivision scenarios, and three sizes of the founding population. As expected, the lesser the degree of subdivision, the larger the number of junctions in the chromosome. This is because more subdivision leads to smaller subpopulations in which IBD builds up more quickly, reducing the number of junctions formed. When N_0=20, the differences in the expected number of junctions are quite large. For example, chromosomes in the non-subdivided population are expected to have 40% more junctions than those in the population which splits whenever a population size of $2N_0$ is reached. This implies that on average, ancestral chromosome segments are considerably longer in the subdivided population. The differences persist in the larger populations, but are greatly reduced. To understand the practical importance of these differences, we consider the variance of the number of junctions per Morgan, in the following section.



(a) $N_0 = 20$



(b) $N_0 = 100$



(c) $N_0 = 500$

Figure 3.2: Expected number of junctions per Morgan as function of generation, for exponentially growing populations with varying degrees of subdivision. Founder population sizes are (a) 20, (b) 100, and (c) 500.

Table 3.1: Expected number of junctions per Morgan in a chromosome taken from generation 100, for four subdivision scenarios in populations growing exponentially at 4.7% per generation.

	(N_0)		
Scenario	20	100	500
no subdivision	64.9	91.7	98.3
splits at $8N_0$	62.8	90.9	98.1
splits at $4N_0$	57.3	88.9	97.7
splits at $2N_0$	45.0	83.5	96.4

3.1.2 Variance of number of junctions per Morgan

The variance approximations which were developed in Section 2.1.2 can be applied in a straightforward manner to subdivided populations. The relevant historical population sizes required by the equations are simply those within the subdivision of interest. In this section, we apply both the Poisson approximation and Equation 2.27 to approximate the variance of the number of junctions at generation 100 in the subdivided populations discussed in the previous section. We also obtain simulation based estimates of this variance for the populations whose founding population size was $N_0 = 20$.

Example of growing populations with repeated subdivision

Recall that all the populations considered were growing exponentially at a rate of 4.7% per generation, to achieve a 100 fold increase over 100 generations. Founding population sizes were $N_0 = 20$, $N_0 = 100$ and $N_0 = 500$. We considered a non-subdivided population, and populations in which splits occurred whenever the local population size reached $8N_0$, $4N_0$ or $2N_0$. For a given value of N_0, all 4 populations have the same total population size (summed over subpopulations).

Figure 3.3 shows the distributions estimated from 10,000 simulations of the number of junctions in a chromosome of length one Morgan, randomly selected from the population at generation 100, for each of the four populations for which $N_0 = 20$. All distributions have a

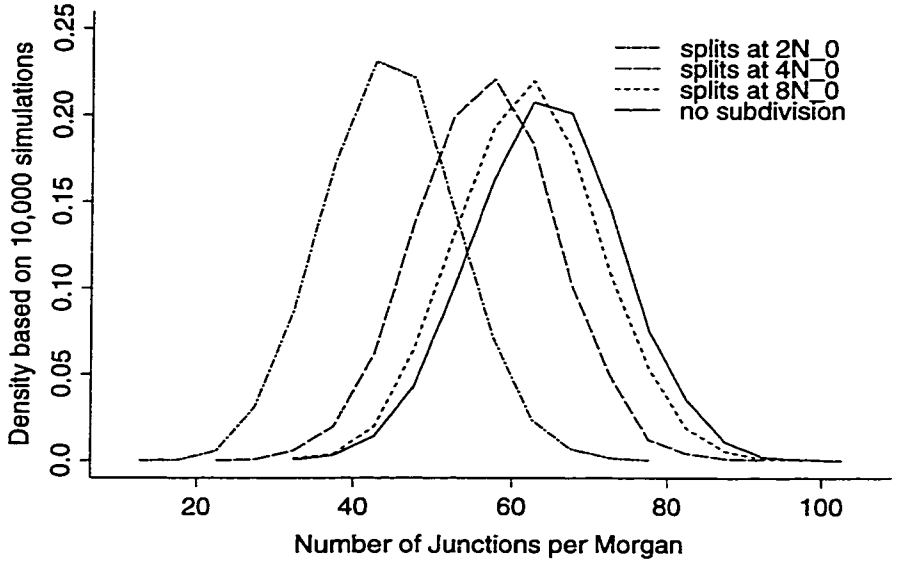


Figure 3.3: Distributions (estimated from 10,000 simulations) of the number of junctions in a randomly selected chromosome from generation 100, for each of 4 growth and subdivision scenarios with $N_0 = 20$.

very large variance. There is considerable overlap between distributions - even between the most subdivided population (splits at $2N_0$) and the non-subdivided population. This shows that while the expected number of junctions per Morgan may be quite different between two populations, the number of junctions actually observed in two populations with different growth patterns could be quite similar by chance.

Table 3.2 quantifies these observations. The table shows the variance of the number of junctions in a chromosome of length one Morgan from generation 100, estimated by (i) simulation, (ii) the Poisson approximation, and (iii) Equation 2.27 of Chapter 2. Simulation based estimates are available only for populations with $N_0 = 20$ and the more subdivided populations with $N_0 = 100$, since simulation of the larger populations is too compuationally demanding. For the populations with $N_0 = 20$, the Poisson approximation badly underestimates the variance. The approximation based on Equation 2.27 is much better, but still an underestimate. This pattern also holds for populations with $N_0 = 100$, and both approxi-

Table 3.2: Variance of the number of junctions in a chromosome randomly selected from generation 100, based on 10,000 simulations, the Poisson approximation, or Equation 2.27.

		no subdiv.	splits at $8N_0$	splits at $4N_0$	splits at $2N_0$
$N_0 = 20$	simulation	86.94	81.85	77.78	68.86
	Poisson	64.93	62.80	57.28	45.00
	Equation 2.27	80.50	76.95	70.93	60.53
$N_0 = 100$	simulation	-	94.07	91.67	86.24
	Poisson	91.74	90.91	88.93	83.45
	Equation 2.27	92.06	91.24	89.34	84.60
$N_0 = 500$	simulation	-	-	-	-
	Poisson	98.29	98.10	97.66	96.36
	Equation 2.27	98.29	98.11	97.66	96.38

mations are quite close to the simulated variance for these populations. For the populations with $N_0 = 500$, the variance approximations are virtually identical, and we hypothesize that the variance is well estimated by either approximation for populations this large.

3.1.3 Expected number of junctions shared per Morgan

As shown in Equation 2.29, the expected number of junctions shared between two chromosomes chosen without replacement from generation t is given by

$$\mathrm{E}[S_t \mid \underline{N}] = \sum_{j=0}^{t-2} \mathrm{E}[n_j] \cdot Pr(Z_t = 2 \mid Y_{j+1} = 1) \ . \tag{3.6}$$

Here n_j is the number of junctions formed in meioses from generation j and $Pr(Z_t = 2 \mid Y_{j+1} = 1)$ is the probability that a junction formed in a meiosis from generation j exists in both sampled chromosomes. To calculate the expected number of junctions shared between two chromosomes from the same subpopulation, we simply apply this equation using the historical population sizes of the subpopulation of interest.

Now consider comparing chromosomes from two different subpopulations at generation

t. If t_s is the time at which the subpopulations split, then

$$\mathrm{E}[S_t \mid \underline{N}] = \sum_{j=0}^{t_s-2} \mathrm{E}[n_j] \cdot Pr(Z_t = 2 \mid Y_{j+1} = 1) \ . \tag{3.7}$$

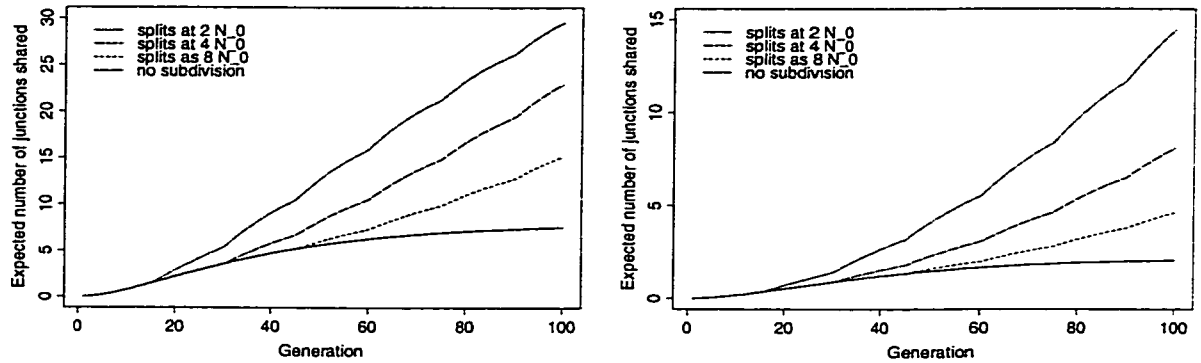
The summation only goes as far as $t_s - 2$, since in order for a junction to be shared, it must have been formed in a meiosis from an ancestor common to both subpopulations. Next we consider the calculation of $Pr(Z_t = 2 \mid Y_{j+1} = 1)$, where $t > t_s$ and $j \leq t_s - 2$. Let l denote the location at which the junction of interest occurred. In the first subpopulation, locus l on the sampled chromosome is a copy of exactly one of the genes present in the founders of that subpopulation. Similarly, locus l in the second chromosome is a copy of exactly one of the genes present in the founders of the second subpopulation. By the "founders of the subpopulation" we mean the individuals in generation t_s who randomly divide into groups to make the subpopulations. Since the subdivision is random, and the subpopulations reproduce according to the rules of random mating, the two subpopulation founder genes represented in the sampled chromosomes can be thought of as two genes chosen without replacement from generation t_s. Therefore

$$\begin{aligned} \mathrm{E}[S_t \mid \underline{N}] &= \sum_{j=0}^{t_s-2} \mathrm{E}[n_j] \cdot Pr(Z_t = 2 \mid Y_{j+1} = 1) \\ &= \sum_{j=0}^{t_s-2} \mathrm{E}[n_j] \cdot Pr(Z_{t_s} = 2 \mid Y_{j+1} = 1) \\ &= \mathrm{E}[S_{t_s} \mid \underline{N}] \ . \end{aligned} \tag{3.8}$$

That is, the expected number of junctions shared between two chromosomes chosen from different subpopulations is equal to the expected number of junctions shared between two chromosomes chosen randomly from the generation in which the split occurred. This result is interesting because it implies that the subpopulation sizes since the subdivision and the number of subdivisions are both irrelevant to the expected number of junctions shared.

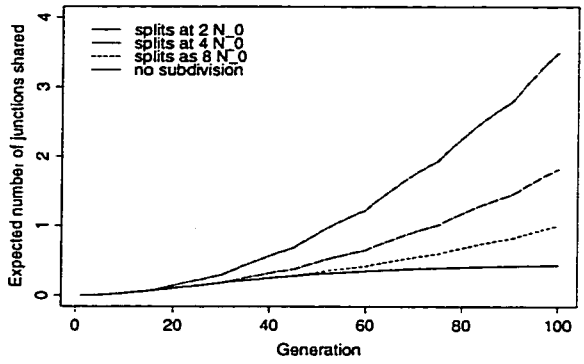
Example of growing populations with repeated subdivision

We return again to our example of populations growing exponentially, with varying degrees of subdivision and different founding population sizes. Figure 3.4 shows the expected num-



(a) $N_0 = 20$

(b) $N_0 = 100$



(c) $N_0 = 500$

Figure 3.4: Expected number of junctions shared per Morgan as function of generation, for exponentially growing populations with varying degrees of subdivision. Founder population sizes are (a) 20, (b) 100, and (c) 500.

ber of junctions shared per Morgan between two chromosomes randomly sampled from the same subpopulation. The three panels correspond to different starting sizes of (a) $N_0 = 20$, (b) $N_0 = 100$ and (c) $N_0 = 500$, and different lines within the same plot correspond to different levels of subdivision. The more subdivided populations always have higher expected junction sharing than less subdivided populations. The relative effect of increasing subdivision is approximately the same regardless of founding population size. However, the absolute magnitude of the expected number of junctions shared is much smaller in the larger populations, which also have a higher expected number of junctions per Morgan overall. We will see an application for the expected number of junctions shared in Chapter 4, when we consider the mean length of an IBD tract between chromosomes.

3.2 Regular Mating Systems

In this section we consider the expected number of junctions in a chromosome randomly selected from a regular mating systems. The systems we consider are repeated sib-mating, repeated double-first-cousin mating, and repeated first-cousin mating. None of these are directly applicable to human populations, but all are interesting theoretical problems of particular interest to plant and animal breeders.

3.2.1 Sib mating

Figure 3.5 shows a system of repeated sib-mating. The founders are two unrelated individuals who have two children. Those children then become the parents of the next generation, and so on. Lines indicate meioses, and filled circles represent individuals. Generation numbers appear down the left side of the figure.

Let h_t be the probability of non-IBD between two genes on homologous chromosomes within an individual in generation t. Similarly, let k_t represent the probability of non-IBD between two genes from different individuals in generation t. Assuming that the founder couple are non-inbred and unrelated, $h_0 = 1$ and $k_0 = 1$. One can show that

$$h_t = k_{t-1} \quad \text{and} \quad k_t = \frac{1}{2}k_{t-1} + \frac{1}{4}h_{t-1} \ , \tag{3.9}$$

and $h_t \to 0$ as $t \to \infty$, and the rate of decay is exponential (see for example [6]).

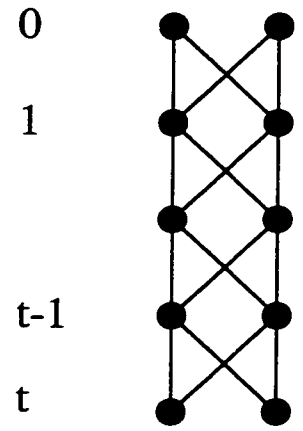


Figure 3.5: Repeated sib-mating.

We now consider the expected number of junctions present in a chromosome of length L, chosen at random from generation t. As we argued previously for the random mating populations,

$$\mathrm{E}[J_t] = \sum_{i=0}^{t-1} \mathrm{E}[n_i] \cdot \mathrm{Pr}_t(i) \ , \tag{3.10}$$

where $\mathrm{Pr}_t(i)$ is the probability that a junction formed in a meiosis from generation i exists in the randomly sampled chromosome from time t, and n_i is the number of junctions formed in meioses from generation i. Consider first the calculation of the probability $\mathrm{Pr}_t(i)$. A junction formed in a meiosis from generation i will be present in exactly one copy in generation $i+1$. Therefore its frequency in generation $i+1$ is exactly 0.25. Since the individuals in generation $i+1$ contribute equally to subsequent generations, and since we are randomly selecting a chromosome from generation t, the 4 genes present at the location of the junction in generation $i+1$ are equally likely to be the ancestor of the same location in the sampled chromosome. Therefore $\mathrm{Pr}_t(i) = 0.25$, for all t greater than i. Consider the calculation of $\mathrm{E}[n_i]$. Now, n_i is the sum of the number of junctions formed in each of the four meioses from generation i. Consider meiosis m from generation i, and let $H_i(m)$ denote the proportion of the chromosome which is not IBD in the person from whom that meiosis originates. If recombinations happen as a Poisson process along the chromosome, this process must have rate one per Morgan, and the number of recombinations in a chromosome of length L has

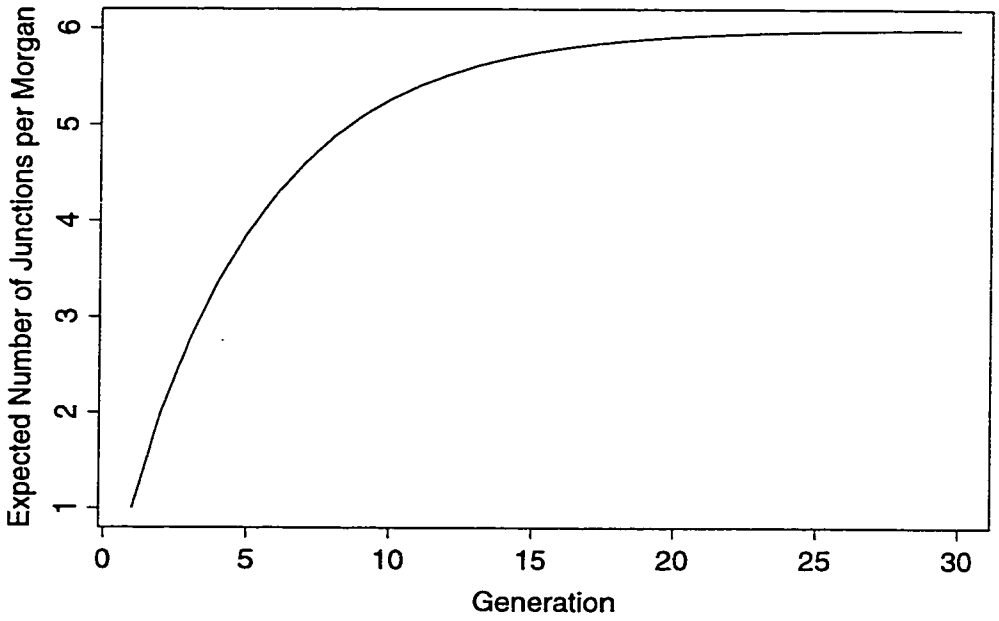


Figure 3.6: Expected number of junctions per Morgan in a chromosome sampled from generation t, repeated sib mating.

a Poisson distribution with mean L. A junction is just a recombination event which occurs in a region of non-IBD, and so conditional on $H_i(m)$, the number of junctions formed in meiosis m has a Poisson distribution with mean $H_i(m)L$. Then

$$\mathrm{E}[n_m] = \mathrm{E}_H[\mathrm{E}[n_m|H_i(m)]] = \mathrm{E}[H_i(m)L] = h_i L \ .$$

Since $n_i = \sum_{m=1}^{4} n_m$, $\mathrm{E}[n_i] = 4h_i L$. Therefore, we can calculate $\mathrm{E}[J_t]$ by

$$\mathrm{E}[J_t] = \sum_{i=0}^{t-1} \mathrm{E}[n_i] \cdot \mathrm{Pr}_t(i) = \sum_{i=0}^{t-1} 4h_i L \cdot 0.25 = \sum_{i=0}^{t-1} h_i L \ . \tag{3.11}$$

Figure 3.6 shows the expected number of junctions per Morgan in a chromosome sampled from generation t for repeated sib-mating. The plot suggests that the expected number of junctions per Morgan converges to six as $t \to \infty$. This can be confirmed algebraically. By Equation 3.9, we have

$$h_{i+1} = \frac{1}{2} h_i + \frac{1}{4} h_{i-1} \ . \tag{3.12}$$

Summing both sides from $i = 1$ to $i = t$, we obtain

$$\begin{aligned}\sum_{i=1}^{t} h_{i+1} &= \frac{1}{2}\sum_{i=1}^{t} h_i + \frac{1}{4}\sum_{i=1}^{t} h_{i-1} \\ \sum_{i=2}^{t+1} h_i &= \frac{1}{2}\sum_{i=1}^{t} h_i + \frac{1}{4}\sum_{i=0}^{t-1} h_i \ .\end{aligned}$$

Since $h_0 = h_1 = 1$, this is equivalent to

$$\sum_{i=0}^{t-1} h_i + h_t + h_{t+1} - 2 = \frac{1}{2}\left(\sum_{i=0}^{t-1} h_i + h_t - 1\right) + \frac{1}{4}\sum_{i=0}^{t-1} h_i \ .$$

Solving for $\sum_{i=0}^{t-1} h_i$, we obtain

$$\sum_{i=0}^{t-1} h_i = 4\left(-\frac{1}{2}h_t - h_{t+1} + \frac{3}{2}\right) \ . \tag{3.13}$$

Taking the limit of Equation 3.13 as $t \to \infty$ gives

$$\lim_{t\to\infty} \sum_{i=0}^{t-1} h_i = 6 \ , \tag{3.14}$$

since $h_t \to 0$ as $t \to \infty$. Therefore $\mathrm{E}[J_t] \to 6\,L$ as $t \to \infty$.

This result is identical to that obtained by Fisher [9] using matrices of mating types. Since $\mathrm{E}[J_t]$ is finite, J_t is finite with probability one. Since we know that each point on the chromosome is eventually fixed, all junctions are shared when the population is fixed. Thus we expect a total of seven segments per Morgan in a chromosome, once the population has become fixed. The segments are therefore quite long, having mean length at least 14.3 cM.

3.2.2 Double-first-cousin mating

Figure 3.7 shows a system of repeated double-first-cousin mating The founders are four unrelated individuals in two couples, each of which have two children. Mating pairs are formed by pairing one individual from each sibship. These pairs have two children each, and the process is repeated. The couple who will give rise to generation three are double-first-cousins, since they are first-cousins through both their parents. The notation is as in Figure 3.5.

Let h_t be the probability of non-IBD between two genes on homologous chromosomes within an individual in generation t. Similarly, let l_t represent the probability of non-IBD

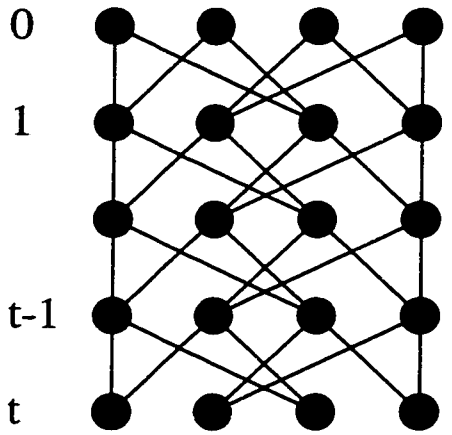


Figure 3.7: Repeated double-first-cousin mating.

between two genes in non-siblings in generation t, and let k_t be the probability of non-IBD between two genes in siblings in generation t. Assuming that the founder couples are non-inbred and unrelated, $h_0 = 1$, $l_0 = 1$ and $k_0 = 1$. One can show that

$$h_t = l_{t-1},\ l_t = \frac{1}{2}k_{t-1} + \frac{1}{2}l_{t-1} \text{ and } k_t = \frac{1}{2}l_{t-1} + \frac{1}{4}h_{t-1}\ , \tag{3.15}$$

and $h_t \to 0$ as $t \to \infty$, and the rate of decay is exponential (see for example [6]).

We now wish to calculate

$$\mathrm{E}[J_t] = \sum_{i=0}^{t-1} \mathrm{E}[n_i] \cdot \mathrm{Pr}_t(i)\ , \tag{3.16}$$

where $\mathrm{Pr}_t(i)$ is the probability that a junction formed in a meiosis from generation i exists in the randomly sampled chromosome from time t, and n_i is the number of junctions formed in meioses from generation i. Using reasoning analogous to that explained in the previous section, we obtain $\mathrm{Pr}_t(i) = \frac{1}{8}$ (since there are eight chromosomes in each generation), and $\mathrm{E}[n_i] = 8h_iL$. Thus we again obtain

$$\mathrm{E}[J_t] = \sum_{i=0}^{t-1} \mathrm{E}[n_i] \cdot \mathrm{Pr}_t(i) = \sum_{i=0}^{t-1} h_i L\ . \tag{3.17}$$

Figure 3.8 shows the expected number of junctions per Morgan in a chromosome sampled from generation t for repeated double first cousin mating. The plot suggests that the

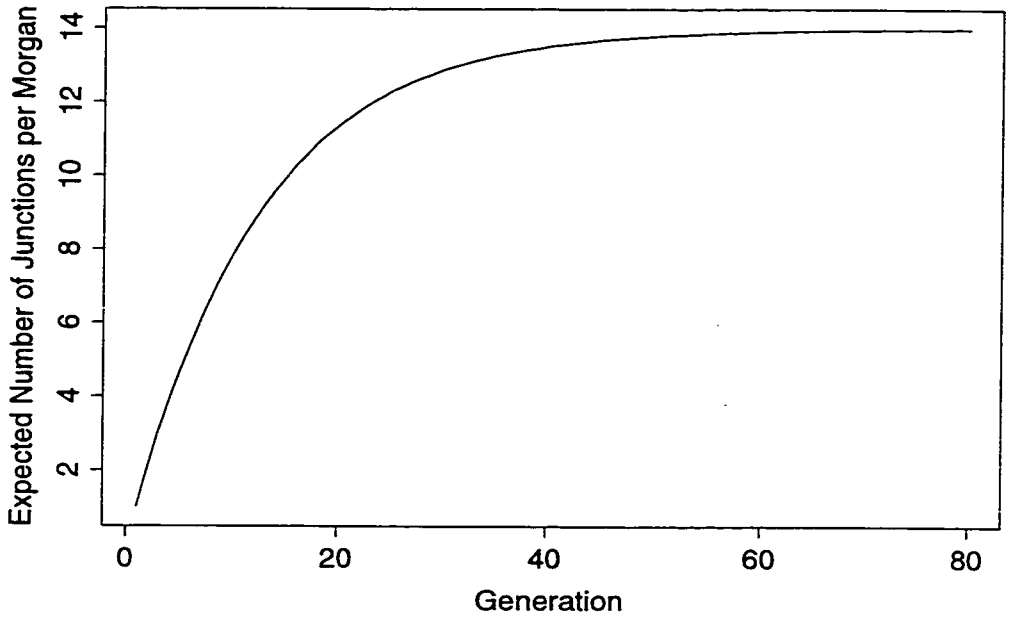


Figure 3.8: Expected number of junctions per Morgan in a chromosome sampled from generation t, repeated double first cousin mating.

expected number of junctions per Morgan converges to fourteen as $t \to \infty$. This can be confirmed algebraicly. By Equation 3.15, we have

$$h_{i+1} = \frac{1}{8}h_{i-2} + \frac{1}{4}h_{i-1} + \frac{1}{2}h_i \ . \tag{3.18}$$

Summing both sides from $i = 2$ to $i = t + 1$, we obtain

$$\begin{aligned} \sum_{i=2}^{t+1} h_{i+1} &= \frac{1}{8}\sum_{i=2}^{t+1} h_{i-2} + \frac{1}{4}\sum_{i=2}^{t+1} h_{i-1} + \frac{1}{2}\sum_{i=2}^{t+1} h_i \\ \sum_{i=3}^{t+2} h_i &= \frac{1}{8}\sum_{i=0}^{t-1} h_i + \frac{1}{4}\sum_{i=1}^{t} h_i + \frac{1}{2}\sum_{i=2}^{t+1} h_i \ . \end{aligned}$$

Since $h_0 = h_1 = h_2 = 1$, this is equivalent to

$$\sum_{i=0}^{t-1} h_i + h_t + h_{t+1} + h_{t+2} - 3 = \frac{1}{8}\sum_{i=0}^{t-1} h_i + \frac{1}{4}\left(\sum_{i=0}^{t-1} h_i + h_t - 1\right) + \frac{1}{2}\left(\sum_{i=0}^{t-1} h_i + h_t + h_{t+1} - 2\right) \ .$$

Solving for $\sum_{i=0}^{t-1} h_i$, we obtain

$$\sum_{i=0}^{t-1} h_i = 8\left(-\frac{1}{4}h_t - \frac{1}{2}h_{t+1} - h_{t+2} + \frac{7}{4}\right) \ . \tag{3.19}$$

Taking the limit of Equation 3.19 as $t \to \infty$ gives

$$\lim_{t\to\infty} \sum_{i=0}^{t-1} h_i = 14 \ , \tag{3.20}$$

since $h_t \to 0$ as $t \to \infty$. Therefore $\mathrm{E}[J_t] \to 14\,L$ as $t \to \infty$.

Since $\mathrm{E}[J_t]$ is finite, J_t is finite with probability one. Since we know that each point on the chromosome is eventually fixed, all junctions are shared when the population is fixed. Thus we expect a total of fifteen segments per Morgan in a chromosome, once the population has become fixed. The segments are likely to be much shorter than in the sib-mating population, having mean length of at least 6.7 cM.

3.2.3 First-cousin mating

Figure 3.9 shows a system of repeated first-cousin mating starting with ten unrelated individuals and continuing for five generations. In this diagram, solid circles represent individuals, and the intersections of lines are therefore matings. This mating system is more interesting

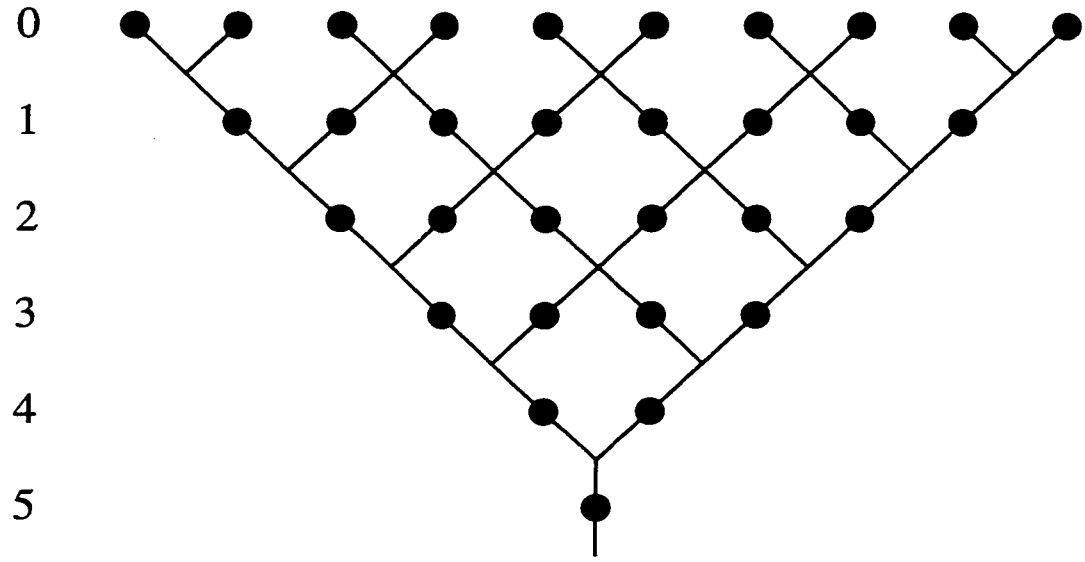


Figure 3.9: Repeated first-cousin mating.

than sib-mating and double-first-cousin mating because it may be approximately representative of some human populations. The linear increase in the number of ancestors looking backward in time may reflect that seen in some human populations, and such a system can be maintained over many generations. In addition, in any given generation, not all individuals contribute equally to the next generation. This is more realistic than the other systems in which all individuals contribute equally.

Let $h_t^{(0)}$ denote the probability of non-IBD for two genes randomly sampled from siblings in generation t. Similarly, let $h_t^{(j)}$ denote the probability of non-IBD for two genes randomly sampled from cousins of degree j in generation t. Finally, let h_t denote the probability of non-IBD for two genes sampled from the same individual in generation t. Since all matings happen between individuals who are first cousins, $h_t = h_{t-1}^{(1)}$. If the founding individuals are assumed to be non-inbred and unrelated, the following initial conditions hold:

$$\begin{aligned} h_0 = h_0^{(j)} &= 1, \quad \text{for all } j = 0, 1, 2, \ldots \\ h_t^{(j)} &= 1, \quad \text{for all } j \geq t, \\ h_1^{(0)} &= 0.75\ . \end{aligned} \tag{3.21}$$

For all $j < t$, the following relationships hold (see, for example [17]):

$$h_t^{(j)} = 0.25 \cdot h_{t-1}^{(j-1)} + 0.5 \cdot h_{t-1}^{(j)} + 0.25 \cdot h_{t-1}^{(j+1)} \quad \text{if } j \geq 1 \tag{3.22}$$

and

$$h_t^{(0)} = 0.5 \cdot h_{t-1}^{(1)} + 0.25 \cdot h_{t-2}^{(1)} \quad \text{if } t \geq 2 \ . \tag{3.23}$$

It may help to think of the $h_t^{(j)}$ as the elements of a matrix, where the rows are indexed by t (generation) and the columns are indexed by j (cousin type). All upper diagonal elements of the matrix are 1, and the lower elements are given by the recursion relations above. It has been shown that $h_t^{(j)} \to 0$ as $t \to \infty \ \forall j$ [17], [24], but the rate of decay is proportional to $\frac{1}{\sqrt{t}}$ [24] – much smaller than for sib mating or double first cousin mating.

We would like to calculate the expected number of junctions per Morgan in a chromosome chosen from an individual who is the result of t generations of first-cousin mating. We use the equation

$$\mathrm{E}[J_t] = \sum_{i=0}^{t-1} \sum_{m=1}^{m_i} \mathrm{E}[n_m] \cdot p_t(m,i) \ , \tag{3.24}$$

where m_i is the number of meioses from generation i, $p_t(m,i)$ is the probability that a junction formed in meiosis m from generation i exists in the chromosome sampled from generation t, and n_m is the number of junctions formed in meiosis m from generation i. This is similar to the approach used for sib mating, and double-first-cousin mating, only we sum over the individual meioses in each generation. Since all individuals in generation i have the same probability of non-IBD h_i, $\mathrm{E}[n_m] = h_i L$ for all $m = 1, \ldots, m_i$. Therefore

$$\mathrm{E}[J_t] = \sum_{i=0}^{t-1} h_i L \cdot \sum_{m=1}^{m_i} p_t(m,i) \ , \tag{3.25}$$

Now consider the probabilities $p_t(m,i)$, $m = 1, \ldots, m_i$. The probability $p_t(m,i)$ can be thought of as the probability that at a particular locus, the product of meiosis m from generation i is IBD to the gene at that locus in the chromosome sampled at time t. The sum of these probabilities over all the meioses from generation i must be equal to one, since one of those products of meioses must be the ancestor of the chromosome sampled from generation t, at the locus of interest. We therefore obtain the familiar equation

$$\mathrm{E}[J_t] = \sum_{i=0}^{t-1} h_i L \ . \tag{3.26}$$

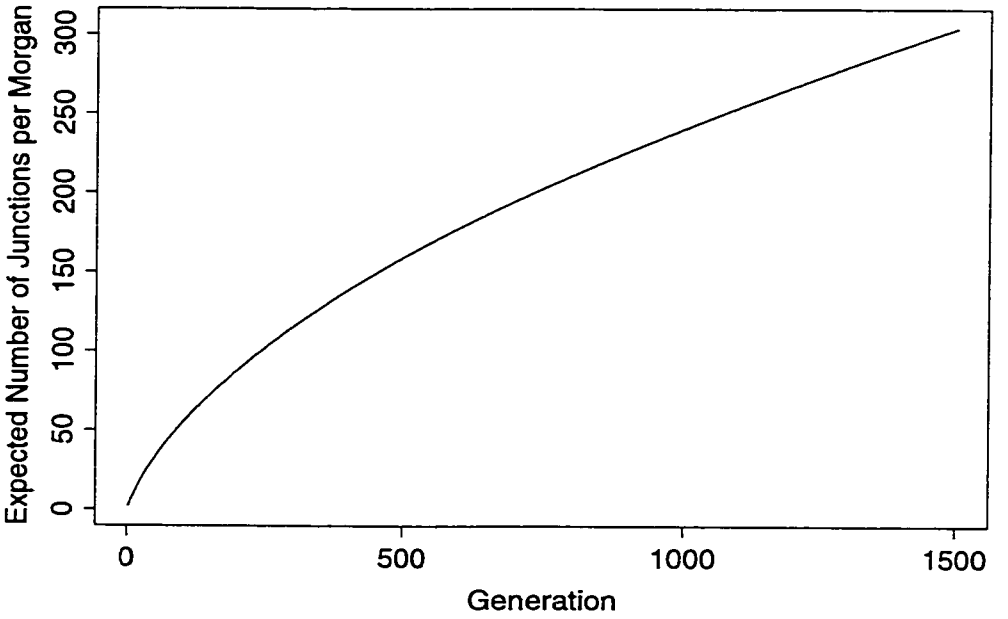


Figure 3.10: Expected number of junctions per Morgan in a chromosome sampled from generation t, repeated first cousin mating.

Figure 3.10 shows the expected number of junctions per Morgan in a chromosome sampled from generation t for repeated first cousin mating. Even after a relatively long period of inbreeding, the expected number of junctions does not appear to converge. In fact, the expected number of junctions per Morgan diverges, as we now prove.

E[J_t] *diverges*

First rewrite Equation 3.26 as

$$\mathrm{E}[J_n] = \left(1 + \sum_{t=1}^{n-1} h_{t-1}^{(1)}\right) L = \left(2 + \sum_{t=1}^{n-2} h_t^{(1)}\right) L \ , \tag{3.27}$$

using the fact that $h_0 = 1$ and $h_t = h_{t-1}^{(1)}$ for all $t > 1$. We show that

$$\sum_{t=1}^{n-2} h_t^{(1)} \to \infty \quad \text{as} \quad t \to \infty \ . \tag{3.28}$$

Let j index cousin type. Then for $j = 1, 2, 3 \ldots$,

$$\begin{aligned} \sum_{t=1}^{n-2} h_t^{(j)} &= 0.25 \cdot \sum_{t=1}^{n-2} h_{t-1}^{(j-1)} + 0.5 \cdot \sum_{t=1}^{n-2} h_{t-1}^{(j)} + 0.25 \cdot \sum_{t=1}^{n-2} h_{t-1}^{(j+1)} \\ &= 0.25 \cdot \left(\sum_{t=1}^{n-3} h_t^{(j-1)} + 1\right) + 0.5 \cdot \left(\sum_{t=1}^{n-3} h_t^{(j)} + 1\right) + 0.25 \cdot \left(\sum_{t=1}^{n-3} h_t^{(j+1)} + 1\right), \end{aligned}$$

so that

$$\sum_{t=1}^{n-2} h_t^{(j)} = 0.25 \cdot \sum_{t=1}^{n-3} h_t^{(j-1)} + 0.5 \cdot \sum_{t=1}^{n-3} h_t^{(j)} + 0.25 \cdot \sum_{t=1}^{n-3} h_t^{(j+1)} + 1. \tag{3.29}$$

For $j = 0$,

$$\begin{aligned} \sum_{t=1}^{n-2} h_t^{(0)} &= 0.75 + 0.5 \cdot \sum_{t=2}^{n-2} h_{t-1}^{(1)} + 0.25 \cdot \sum_{t=2}^{n-2} h_{t-2}^{(1)} \\ &= 0.75 + 0.5 \cdot \sum_{t=1}^{n-3} h_t^{(1)} + 0.25 \left(\sum_{t=1}^{n-4} h_t^{(1)} + 1\right) \end{aligned}$$

so that

$$\sum_{t=1}^{n-2} h_t^{(0)} = 1 + 0.5 \cdot \sum_{t=1}^{n-3} h_t^{(1)} + 0.25 \cdot \sum_{t=1}^{n-4} h_t^{(1)} \tag{3.30}$$

Since $0 \le h_t^{(j)} \le 1$, $\sum_{t=1}^{n-2} h_t^{(j)}$ is non-decreasing with n, for all j. This means that as $n \to \infty$, either $\sum_{t=1}^{n-2} h_t^{(j)} \to \alpha_j$ where α_j is some finite positive number, or $\sum_{t=1}^{n-2} h_t^{(j)} \to \infty$.

Equations 3.29 and 3.30 show that if any one of the sums indexed by j diverges, then they all must. We now show that $\sum_{t=1}^{n-2} h_t^{(1)}$ diverges, by contradiction. First we note that

$$h_t^{(0)} \leq h_t^{(1)} \leq h_t^{(2)} \leq h_t^{(3)} \ldots \tag{3.31}$$

for all t. We prove Equation 3.31 by induction on t. Note that the initial conditions imply the truth of Equation 3.31 for $t = 0, 1, 2$. Now, assuming that Equation 3.31 holds for $t \leq n$, we will show that it holds for $t = n + 1$. For $j = 1, 2, 3, \ldots$, we have

$$\begin{aligned} h_{n+1}^{(h+1)} &= 0.25 \cdot h_n^{(j)} + 0.5 \cdot h_n^{(j+1)} + 0.25 \cdot h_n^{(j+2)} \\ &\geq 0.25 \cdot h_n^{(j-1)} + 0.5 \cdot h_n^{(j)} + 0.25 \cdot h_n^{(j+1)} \\ &= h_{n+1}^{(j)} \end{aligned}$$

So for all $j \geq 1$, we have shown $h_{n+1}^{(j)} \leq h_{n+1}^{(j+1)}$. It remains only to show that $h_{n+1}^{(0)} \leq h_{n+1}^{(1)}$. Consider

$$\begin{aligned} h_{n+1}^{(1)} - h_{n+1}^{(0)} &= 0.25 \cdot h_n^{(0)} + 0.5 \cdot h_n^{(1)} + 0.25 \cdot h_n^{(2)} - 0.5 \cdot h_n^{(1)} - 0.25 \cdot h_{n-1}^{(1)} \\ &= 0.25 \cdot h_n^{(0)} + 0.25 \cdot \left(h_n^{(2)} - h_{n-1}^{(1)}\right) \\ &= 0.25 \cdot h_n^{(0)} + 0.25 \cdot \left(0.25 \cdot h_{n-1}^{(1)} + 0.5 \cdot h_{n-1}^{(2)} + 0.25 \cdot h_{n-1}^{(3)} - h_{n-1}^{(1)}\right) \\ &= 0.25 \cdot h_n^{(0)} + 0.25 \cdot \left(0.25 \cdot (h_{n-1}^{(3)} - h_{n-1}^{(2)}) + 0.75 \cdot (h_{n-1}^{(2)} - h_{n-1}^{(1)})\right) \\ &\geq 0 \,, \end{aligned}$$

since $h_{n-1}^{(3)} \geq h_{n-1}^{(2)}$ and $h_{n-1}^{(2)} \geq h_{n-1}^{(1)}$, by assumption. Thus $h_{n+1}^{(0)} \leq h_{n+1}^{(0)}$. So by induction on t, we have shown that

$$h_t^{(j)} \leq h_t^{(j+1)} \quad \forall j = 0, 1, 2, \ldots \tag{3.32}$$

for all t.

Suppose that

$$\lim_{n \to \infty} \sum_{t=1}^{n-2} h_t^{(j)} = \alpha_j \quad j = 0, 1, 2, \ldots \tag{3.33}$$

where α_j is some positive finite number. Then Equation 3.32 implies that

$$\alpha_j = \lim_{n \to \infty} \sum_{t=1}^{n-2} h_t^{(j)} \leq \lim_{n \to \infty} \sum_{t=1}^{n-2} h_t^{(j)} = \alpha_{j+1} \quad \forall j, \tag{3.34}$$

so that

$$\alpha_j \leq \alpha_{j+1} \quad \text{for all } j. \tag{3.35}$$

However, Equation 3.29 shows that

$$\alpha_j = 0.25 \cdot \alpha_{j-1} + 0.5 \cdot \alpha_j + 0.25 \cdot \alpha_{j+1} + 1, \quad j \geq 1,$$

so

$$\begin{aligned} 0.25 \cdot \alpha_{j+1} - 0.25 \cdot \alpha_j &= 0.25 \cdot \alpha_j - 0.25 \cdot \alpha_{j-1} - 1 \\ \alpha_{j+1} - \alpha_j &= \alpha_j - \alpha_{j-1} - 4 \\ &= (\alpha_{j-1} - \alpha_{j-2} - 4) - 4 \\ &= \ldots \\ &= \alpha_1 - \alpha_0 - 4j \ . \end{aligned}$$

Since α_1 and α_0 are positive finite numbers, this shows that for large enough j, $\alpha_{j+1} - \alpha_j < 0$, which implies

$$\alpha_{j+1} < \alpha_j \ . \tag{3.36}$$

Notice that Equations 3.35 and 3.36 contradict one another. Therefore supposition 3.33 must be false, and thus at least one of the sums $\sum_{t=1}^{\infty} h_t^{(j)}$ must diverge. However, Equations 3.29 and 3.30 showed that if one of the sums diverges, they all must. Therefore

$$\sum_{t=1}^{n-2} h_t^{(1)} \to \infty \quad \text{as} \quad t \to \infty \ , \tag{3.37}$$

and so

$$\mathrm{E}[J_n] \to \infty \quad \text{as} \quad n \to \infty. \tag{3.38}$$

Thus if we think of an infinitely large population reproducing by first cousin mating and progressing towards fixation, we expect the chromosomes to be broken into infinitesimally small pieces relative to the founder population.

3.3 Discussion

In this chapter, we showed how the results of the previous chapter can be simply applied to subdivided populations. We considered some example populations, and examined the effects

of different amounts of subdivision on the expected number of junctions in a chromosome and the expected number of junctions shared between two chromosomes. We found that subdivision can substantially reduce the expected number of junctions in a chromosome, particularly in small populations. As we saw in Chapter 2, the variance of the number of junctions in a chromosome is very large. We also showed that the greater the extent of subdivision, the more junctions are expected to be shared between chromosomes within a subpopulation. The relative effect of subdivision on junction sharing is similar regardless of population size, but sharing is greatly reduced in large populations.

In the second section, we presented some results for regular mating systems. We confirmed a known result for the limiting number of junctions in sib-mating, and presented an original derivation of a similar result for double-first-cousin mating. We also showed that the expected number of junctions in first-cousin mating diverges as t goes to infinity. This is interesting because it suggests that the ancestral segments in the fixed chromosome will be infinitesimally small.

Chapter 4

THE EFFECTS OF POPULATION SIZE, GROWTH AND SUBDIVISION ON THE LENGTHS OF IBD REGIONS

In addition to studying the expected number of ancestral segments in a chromosome, it is also of interest to consider the tracts of IBD between two chromosomes sampled from the population. In this chapter we discuss an approach to modelling the length of an IBD tract, using the results from chapters 2 and 3, and results from the literature. We assess the accuracy of our model, and use it to investigate the effects of population growth and subdivision on the mean length of tracts of IBD.

Figure 4.1 depicts two chromosomes sampled from a population some time after founding. Different patterns represent different ancestral chromosomes. Tracts of IBD and non-IBD are indicated above the chromosomes by black (non-IBD) and white (IBD) bars. Note that a single tract of IBD or non-IBD is made up of a variable number of segments, where a segment is defined as the piece of the chromosome between two neighbouring junctions. In order to precisely describe the tracts of IBD and non-IBD, we classify each junction according to the IBD state between the pair of chromosomes on either side of the junction. Thus there are four types of junction; those non-IBD on both sides, those non-IBD on the left and IBD on the right, those IBD on the left and non-IBD on the right, and those which are IBD on both sides. These junction types will be referred to as types U, V, T and S in what follows. Type S junctions are the shared junctions of the previous chapter, and junction types T, U and V are unshared junctions, which we previously denoted as a group with the letter U. Note that a tract of IBD starts with a type V junction, followed by a variable number K (which could be zero) of type S junctions, and ends with a type T junction. Similarly, a tract of non-IBD starts with a type T junction, continues with a variable number M (which could be zero) of type U junctions, and ends with a type V junction. The process can be thought of in two components; the first component is the sequence of

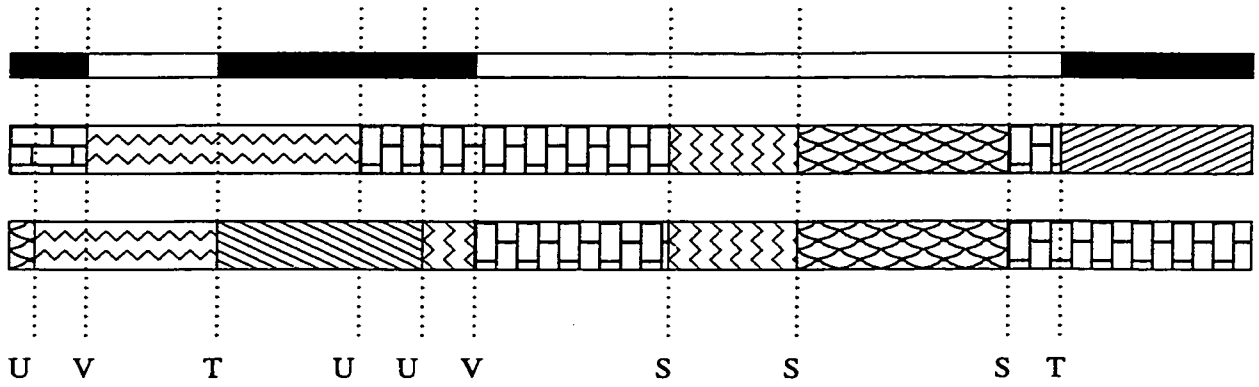


Figure 4.1: Two chromosomes sampled from a population, some time after founding. Different shades represent different ancestral chromosomes.

junction types along the pair of chromosomes, and the second component is the lengths of the segments in between each adjacent pair of junctions. In order to model the distribution of the length of an IBD tract, we must make an assumption first about the occurrence of the different junction types along the chromosome, which tells us about the distribution of K, and second about the distributions of segment lengths.

In order to assess the models we describe in the next section, we conducted some simulation studies. The details of simulation in a random mating population are described in more detail in Chapter 5. Conditional on population sizes over time, random mating populations were simulated. For each generation, two chromosomes were randomly selected from the population. The two were compared, and the tracts of IBD were examined. An IBD tract was then chosen at random, and its length and the number K of internal type S junctions were recorded. The mean and variance of the length of an IBD tract, and the mean number of type S junctions were obtained by repeated simulation. For a select few generations, the entire simulation distribution of K was recorded. These simulation results estimate the true distributions, which we try to model in the next section.

4.1 Modelling the length of an IBD tract

4.1.1 Model A: Independence of Junction Types and IID Exponential Segments

We first used a very simplistic model, which we present here because it helps to motivate the better models. To start, we consider only two classes of junctions - type S and not type S. We denote these classes by S and $\tilde{\text{S}}$. We assume that these two classes appear independently along the chromosome. That is, the probability of a particular junction being of class S does not depend on the classes of neighbouring junctions, only the overall probability of that class of junction, which we denote π_S. Under this assumption K has distribution

$$Pr(K = k) = \begin{cases} 1 - \pi_S & k = 0 \\ \pi_S^k \cdot (1 - \pi_S) & k \geq 1 \end{cases} \tag{4.1}$$

and so $K + 1$, which is the number of segments in a randomly chosen IBD tract, has a geometric distribution with mean $\frac{1}{1-\pi_S}$.

In order to model the lengths of the individual segments, we assumed that junctions occur along the pair of chromosomes (regardless of class) as a Poisson process. The expected number of junctions in a single Morgan of the pair of chromosomes is given by

$$\mathrm{E}[J_t^c|\underline{N}] = 2 \cdot \mathrm{E}[J_t|\underline{N}] - \mathrm{E}[S_t|\underline{N}], \tag{4.2}$$

where J_t and S_t denote the number of junctions in a chromosome randomly sampled from generation t and the number of junctions shared between two chromosomes randomly sampled from generation t, respectively. The expectations of these quantities were developed in chapters 2 and 3. This equation reflects that the number of junctions present in one or both chromosomes is simply the number expected in the first one, plus the number expected in the second one, minus those which are present in both. The Poisson process therefore has rate $\mathrm{E}[J_t^c|\underline{N}]$, and so the segment lengths have an exponential distribution with mean $1/\mathrm{E}[J_t^c|\underline{N}]$.

Let L_I denote the length of a randomly selected IBD tract. Then one can write

$$L_I = \sum_{i=1}^{K+1} X_i \tag{4.3}$$

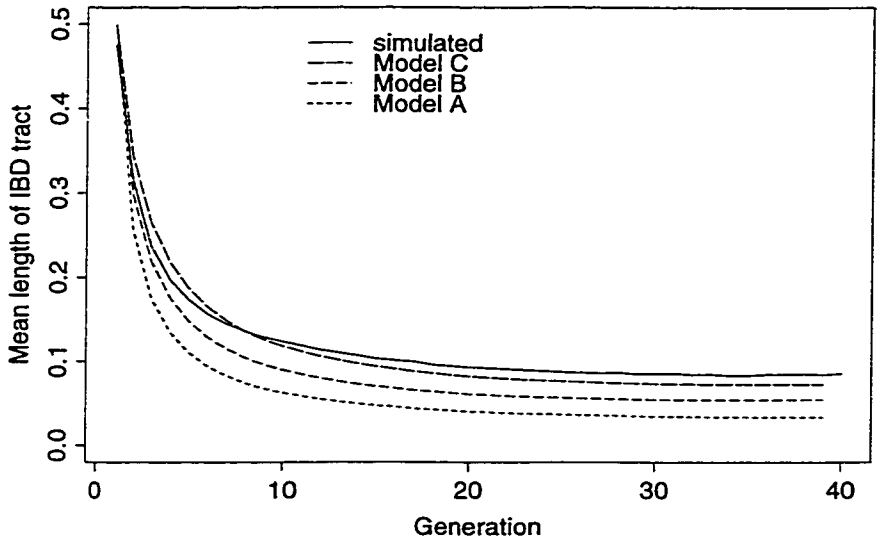
where X_i is the length of ith segment of the tract. Under the assumptions of model A, L_I is the sum of a geometrically distributed number of independent and exponentially distributed random variables. Let λ denote the mean of the exponential distribution, and let $1/p$ denote the mean of the geometric distribution. Now consider the moment generating function of L_I:

$$\begin{aligned}
\mathrm{E}[e^{tL_I}] &= \mathrm{E}_K[\mathrm{E}[e^{tL_I}|K+1=k]] \\
&= \mathrm{E}_K[\prod_{i=1}^{k}\mathrm{E}[e^{tX_i}]] \\
&= \mathrm{E}_k[(1-t\lambda)^{-k}] \\
&= \sum_{k=1}^{\infty}(1-t\lambda)^{-k}\cdot p(1-p)^{k-1} \\
&= \frac{p}{1-t\lambda}\sum_{k=1}^{\infty}\frac{(1-p)^{k-1}}{(1-t\lambda)^{k-1}} \\
&= \frac{p}{1-t\lambda}\cdot\frac{1-t\lambda}{p-t\lambda}\cdot\sum_{k=1}^{\infty}\left(1-\frac{(1-p)}{(1-t\lambda)}\right)\cdot\frac{(1-p)^{k-1}}{(1-t\lambda)^{k-1}} \\
&= \frac{p}{p-t\lambda} \\
&= \left(1-\frac{t\lambda}{p}\right)^{-1} \qquad (4.4)
\end{aligned}$$

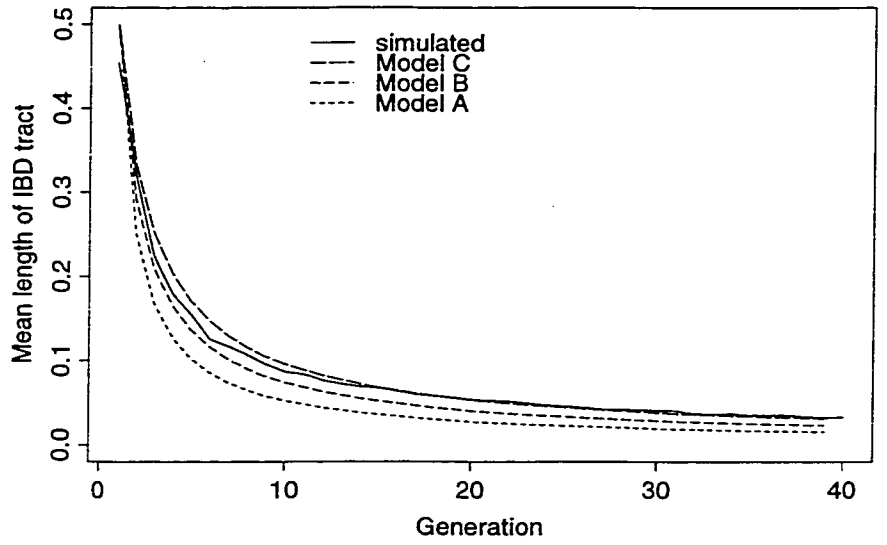
This is the same as the moment generating function for an exponential distribution with mean λ/p. Under the assumptions of model A $\lambda = 1/\mathrm{E}[J_t^c|\underline{N}]$ and $p = 1-\pi_S$, so L_I has an exponential distribution with mean $((1-\pi_S)\cdot\mathrm{E}[J_t^c|\underline{N}])^{-1}$. We approximate π_S by $\mathrm{E}[S_t|\underline{N}]/\mathrm{E}[J_t^c|\underline{N}]$. The performance of this approximation is discussed in Section 4.1.2.

Assessing model fit

Figure 4.2 shows the average length of an IBD tract (based on simulation) and the expected length of an IBD tract based on Model A, for populations of size (a) N=20 and (b) N=100. In both cases, the mean length is badly underestimated by Model A. One possible explanation for this problem is that the distribution of K is not well approximated by Equation 4.1. Table 4.1 shows the empirical distributions of K for a population of size $N = 20$ at generations 5, 10 and 20, based on 100,000 simulations, compared to the predicted distribution under Model A. Table 4.2 shows the same distributions, for a population size of $N = 100$,



(a) N=20



(b) N=100

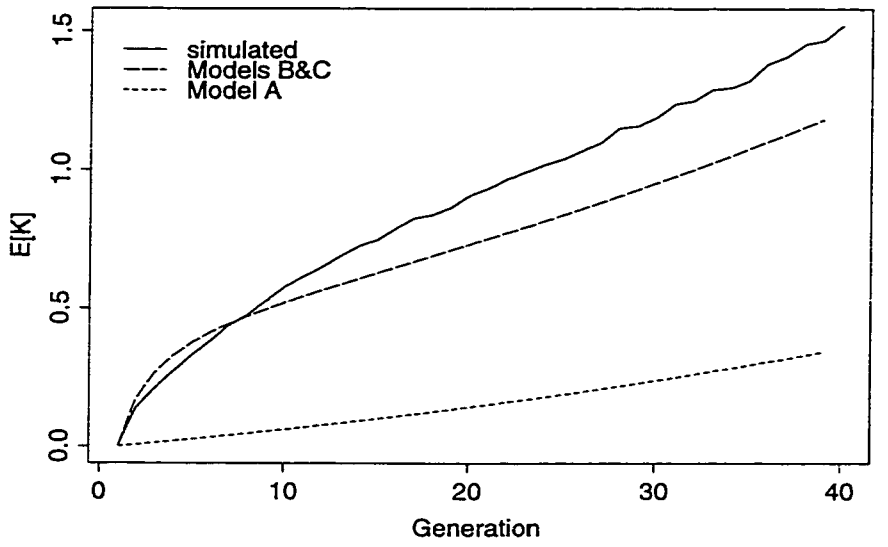
Figure 4.2: Mean length of an IBD tract as estimated by simulation and by models A, B and C, for populations of size (a) N=20, and (b) N=100.

Table 4.1: Empirical distribution of K compared to that predicted under Model A, for N=20, t=5, 10 and 20.

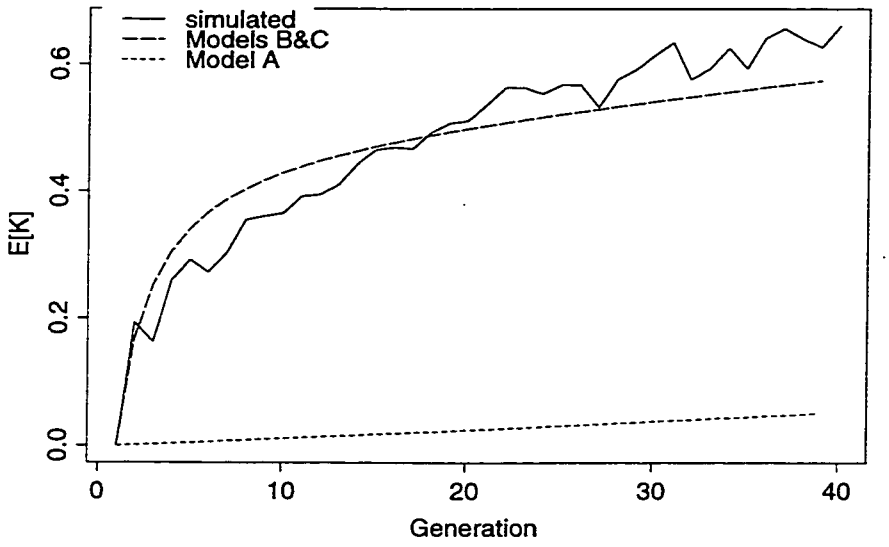
		K							
		0	1	2	3	4	5	>5	mean
t=5	Empirical Dist.	0.819	0.110	0.038	0.015	0.008	0.004	0.006	0.329
	Model A	0.974	0.025	0.001	0	0	0	0	0.026
t=10	Empirical Dist.	0.743	0.143	0.051	0.024	0.013	0.008	0.018	0.578
	Model A	0.942	0.055	0.003	0	0	0	0	0.062
t=20	Empirical Dist.	0.676	0.161	0.066	0.032	0.019	0.011	0.034	0.908
	Model A	0.875	0.109	0.014	0.002	0	0	0	0.143

based on 10,000 simulations. For both populations, model A overestimates the proportion of IBD tracts which have no type S junctions (those that are V-T), and underestimates the proportion which have one or more type S junctions. This results in a severe underestimation of the mean of K, and is one reason why the estimate of mean tract length based on model A performs so poorly. Figure 4.3 demonstrates that this underestimation of the mean of K happens at all generations.

The empirical distributions of K demonstrate that junctions of different type do not occur independently along the chromosome. Instead, junctions of type S come in clusters. For example, consider a type S junction. The next junction in the pair of chromosomes is more likely to be of type S than if the first junction had been some other type. To understand this, consider the meioses in which a junction that is now type S was formed. Any existing junctions in the near neighbourhood of that new junction will likely be inherited with it, and therefore will exist as type S junctions in the pair of chromosomes. This leads to a clustering of junctions of type S. Junctions of type U would also be expected to cluster, for similar reasons. Model B attempts to model the occurrence of junctions types more appropriately.



(a) N=20



(b) N=100

Figure 4.3: Mean number K of type S junctions in an IBD tract, estimated by simulation and by models A, B&C, for populations of size (a) N=20, and (b) N=100.

Table 4.2: Empirical distribution of K compared to that predicted under Model A, for N=100, t=5, 10 and 20.

		K							
		0	1	2	3	4	5	>5	mean
t=5	Empirical Dist.	0.830	0.107	0.037	0.014	0.003	0.003	0.005	0.292
	Model A	0.995	0.005	0	0	0	0	0	0.005
t=10	Empirical Dist.	0.795	0.130	0.042	0.017	0.006	0.004	0.003	0.366
	Model A	0.982	0.017	0	0	0	0	0	0.018
t=20	Empirical Dist.	0.762	0.139	0.044	0.023	0.011	0.007	0.012	0.512
	Model A	0.976	0.023	0.001	0	0	0	0	0.025

4.1.2 Model B: 1st order Markov dependence of junction types and I.I.D. exponential segments

To better approximate the distribution of K, we think of the sequence of junction types along the chromosome as a first order Markov chain on the states {U,V,T,S}. Then the transition matrix for the chain can be written as

$$P_t = \begin{bmatrix} p_1 & 1-p_1 & 0 & 0 \\ 0 & 0 & p_2 & 1-p_2 \\ \theta_1 & 1-\theta_1 & 0 & 0 \\ 0 & 0 & \theta_2 & 1-\theta_2 \end{bmatrix}, \tag{4.5}$$

where p_1 is the probability of staying in state U, p_2 is the probability of moving from state V to state T, θ_1 is the probability of moving from T to U, and θ_2 is the probability of moving from state S to state T. The zeros in the transition matrix are structural - that is, they represent a fundamental constraint on the process of junction types. For example, if a junction is type U, the immediately following junction cannot be type T or S, because the IBD state to the right of the first junction must be equivalent to the IBD state to the left of the second junction.

The Markov model completely specifies the distribution of K. In terms of the parameters

of the transition matrix,

$$Pr(K=k) = \begin{cases} p_2 & k=0 \\ (1-p_2)(1-\theta_2)^{k-1}\theta_2 & k \geq 1 \end{cases}. \tag{4.6}$$

Then

$$\mathrm{E}[K] = \sum_{k=1}^{\infty} k \cdot (1-p_2)(1-\theta_2)^{k-1}\theta_2 = (1-p_2)\sum_{k=1}^{\infty} k(1-\theta_2)^{k-1}\theta_2 = \frac{1-p_2}{\theta_2}. \tag{4.7}$$

Recall that non-IBD tracts start with a type T junction followed by M type U junctions, and end with a type V junction. The distribution of M is also fully specified by the Markov model. The distribution is

$$Pr(M=m) = \begin{cases} 1-\theta_1 & m=0 \\ \theta_1 p_1^{m-1}(1-p_1) & m \geq 1 \end{cases}., \tag{4.8}$$

and so

$$\mathrm{E}[M] = \frac{\theta_1}{1-p_1}. \tag{4.9}$$

From the transition matrix, one can obtain four equations for the equilibrium probabilities of the junction types, in terms of the parameters $p_1, p_2, \theta_1, \theta_2$. Denote the equilibrium probabilities by π_U, π_V, π_T, and π_S. Then

$$\pi_U p_1 + \pi_T \theta_1 = \pi_U \tag{4.10}$$

$$\pi_U(1-p_1) + \pi_T(1-\theta_1) = \pi_V \tag{4.11}$$

$$\pi_V p_2 + \pi_S \theta_2 = \pi_T \tag{4.12}$$

$$\pi_V(1-p_2) + \pi_S(1-\theta_2) = \pi_S \ . \tag{4.13}$$

These equations can be simplified to give

$$\pi_V = \pi_T \tag{4.14}$$

$$\frac{\pi_U}{\pi_T} = \frac{\theta_1}{1-p_1} = \mathrm{E}[M] \tag{4.15}$$

$$\frac{\pi_S}{\pi_T} = \frac{1-p_2}{\theta_2} = \mathrm{E}[K] \ . \tag{4.16}$$

These equations show that if we can estimate the equilibrium probabilities, we can estimate $\mathrm{E}[M]$ and $\mathrm{E}[K]$ under our Markov model. As discussed in the previous section, we can

estimate π_S by $\mathrm{E}[S_t|\underline{N}]/\mathrm{E}[J_t^c|\underline{N}]$. In 1980, Stam [27] described an approach to calculating the expected number of type V or T junctions between two chromosomes. We denote this quantity by $\mathrm{E}[V_t + T_t|\underline{N}]$, and estimate $\pi_V = \pi_T$ by $0.5 \cdot \mathrm{E}[V_t + T_t|\underline{N}]/\mathrm{E}[J_t^c|\underline{N}]$. In the following paragraph we describe Stam's method.

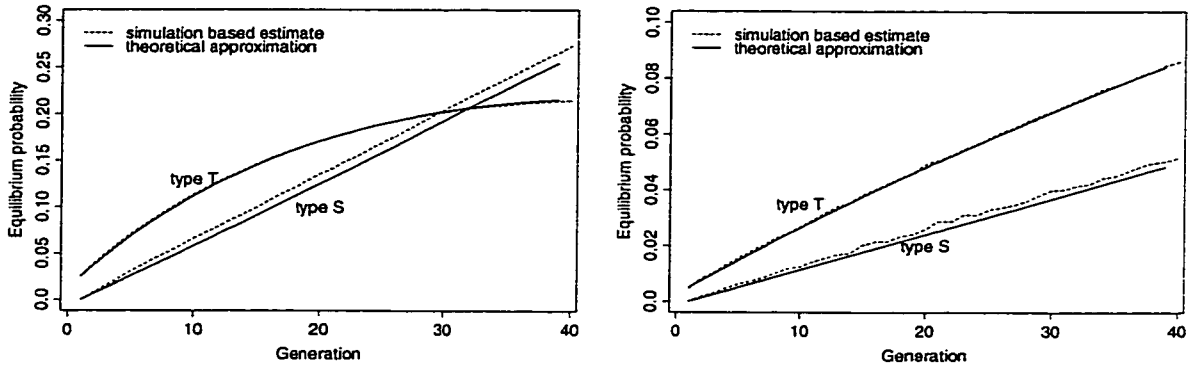
Stam's original result was for a random mating population in which selfing was excluded. We present his equations, simplified for the case where selfing is allowed. In order to calculate $\mathrm{E}[V_t+T_t|\underline{N}]$, we need the probability R_t that three genes chosen randomly without replacement from generation t are all of different ancestral types (all pairs are nonIBD). Now

$$R_{t+1} = \frac{2N_t - 1}{2N_t} \cdot \frac{2N_t - 2}{2N_t} \cdot R_t = \left(1 - \frac{1}{2N_t}\right)\left(1 - \frac{1}{N_t}\right) R_t. \tag{4.17}$$

This equation reflects the fact that in order for three genes to be of different ancestral types, they must have had three distinct parent genes in the parent generation, and those three genes must themselves have been of different ancestral types. In the founding generation R_0 is assumed to be 1, since founders are non-inbred and unrelated, and so R_t is easily calculated conditional on population sizes over time. Now let Q_t denote the number of V or T junctions existing on one of a pair of two chromosomes of length one Morgan chosen randomly from generation t. Then $\mathrm{E}[V_t + T_t|\underline{N}] = 2\mathrm{E}[Q_t|\underline{N}]$, since there are two chromosomes in the pair, and junctions on either chromosome are counted. Stam argues that

$$\mathrm{E}[Q_{t+1}|\underline{N}] = \left(1 - \frac{1}{2N_t}\right)\mathrm{E}[Q_t|\underline{N}] + h_t(\underline{N}) - \left(1 - \frac{1}{N_t}\right) R_t. \tag{4.18}$$

The first term in the equation corresponds to junctions that already existed in the previous generation. In this case, a junction that is of type V or T in two chromosomes in generation t+1 must have been of type V or T in the parent chromosomes, and the same chromosome must not have been chosen twice, an event with probability $1/2N_t$. The second term adds in the expected number of newly formed junctions, and the third term subtracts out the expected number of those which are of type U. Newly formed junctions cannot be of type S, since by definition they exist in only one chromosome of generation t+1. This equation can be applied iteratively to calculate $\mathrm{E}[V_t + T_t|\underline{N}]$, and hence approximate π_V and π_T. Since the equilibrium probabilities sum to one, this result together with our approximation of π_S also gives us an approximation of π_U.



(a) N=20 (b) N=100

Figure 4.4: Estimates of equilibrium probabilities π_S and π_T based on the theoretical approximation and 10,000 simulations, for populations of constant size (a) N=20 and (b) N=100.

Our approximations of the equilibrium probabilities π assume that the expected value of a ratio is equal to the ratio of the expectations. This is only true if the numerator and denominator of the ratio are uncorrelated. We checked the accuracy of this approximation by simulation. Figure 4.4 shows the theoretical estimates of π_S and π_T, as well as the estimates based on 10,000 simulations, for two populations of constant size. For both populations, π_T is well approximated by the ratio of expectations, since the simulation based estimate and the theoretical approximation are nearly indistinguishable. For π_S, the theoretical approximation is an underestimate of the true value for both populations. However, the difference is not large, especially in the larger population.

Recall that

$$L_I = \sum_{i=1}^{K+1} X_i \tag{4.19}$$

where X_i is the length of the ith segment of the tract. We assume as we did in Model A that junctions occur as a Poisson process with rate $\mathrm{E}[J_t^c|\underline{N}]$ along the chromosome, and so the X_i are independent and exponentially distributed with mean $1/\mathrm{E}[J_t^c|\underline{N}]$. Applying

conditional expectation,

$$\begin{aligned} \mathrm{E}[L_I] &= \mathrm{E}_K[\mathrm{E}[\sum_{i=1}^{K+1} X_i \mid K=k]] \\ &= \mathrm{E}_K[\sum_{i=1}^{k+1} \mathrm{E}[X_i|K=k]] \\ &= \mathrm{E}_K[(k+1)\frac{1}{\mathrm{E}[J_t^c|\underline{N}]}] \\ &= \frac{1}{\mathrm{E}[J_t^c|\underline{N}]}(\mathrm{E}[K]+1) \quad . \end{aligned} \tag{4.20}$$

We estimate $\mathrm{E}[K]$ by $\hat{\pi}_S/\hat{\pi}_T$, and thus an estimate of the expected length of an IBD tract is obtained.

Assessing model fit

Figure 4.2 shows the average length of an IBD tract (based on simulation) and the expected length of an IBD tract based on Model B, for populations of size (a) N=20 and (b) N=100. Model B performs better than Model A, although it still underestimates the average length in later generations.

To check the fit of the Markov model, we first must estimate the parameters of the transition matrix by simulation. This was done by randomly selecting a segment from the pair of chromosomes, and classifying it according to the junction types at each end. This was repeated for each iteration of the simulation resulting in counts of the number of each segment type over all simulations. Then, for example, p_1 is estimated by the number of U-U segments, divided by the number of U-U segments plus the number of U-V segments. Table 4.3 shows the estimated transition matrices for generations 5, 10 and 20, for a population of constant size 20. We then used these estimated transition matrices to calculate the expected distribution of K under the Markov model, using Equation 4.7. Table 4.4 shows the empirical distributions of K based on 100,000 simulations, for a population of size $N = 20$ at generations 5, 10 and 20, compared to the predicted distribution under Model B. Model B fits the observed distribution of K remarkably well at generation 10. At generation 5, the model puts too much weight in the tail of the distribution, and therefore overestimates the mean. At generation 20, the model puts too little weight in the tail of the distribution

Table 4.3: Estimates of the transition matrix parameters, based on 100,000 simulations, for $N = 20$ and $t = 5$, 10 or 20.

	t		
	5	10	20
p_1	0.932	0.871	0.771
p_2	0.794	0.745	0.681
θ_1	0.793	0.798	0.708
θ_2	0.490	0.440	0.391

Table 4.4: Empirical distribution of K compared to that predicted under Models B&C, for N=20, t=5, 10 and 20.

		K							
		0	1	2	3	4	5	> 5	mean
t=5	Empirical Dist.	0.819	0.110	0.038	0.015	0.008	0.004	0.006	0.329
	Models B&C	0.794	0.101	0.051	0.025	0.013	0.006	0.010	0.420
t=10	Empirical Dist.	0.743	0.143	0.051	0.024	0.013	0.008	0.018	0.578
	Models B&C	0.745	0.112	0.063	0.035	0.020	0.011	0.014	0.580
t=20	Empirical Dist.	0.676	0.161	0.066	0.032	0.019	0.011	0.034	0.908
	Models B&C	0.681	0.125	0.076	0.046	0.028	0.017	0.027	0.816

Table 4.5: Estimates of the transition matrix parameters, based on 100,000 simulations, for $N = 100$ and $t = 5$, 10 or 20.

	t		
	5	10	20
p_1	0.986	0.972	0.949
p_2	0.782	0.820	0.808
θ_1	0.893	0.921	0.924
θ_2	0.600	0.477	0.419

Table 4.6: Empirical distribution of K compared to that predicted under Models B&C, for N=100, t=5, 10 and 20.

		K							
		0	1	2	3	4	5	>5	mean
t=5	Empirical Dist.	0.830	0.107	0.037	0.014	0.003	0.003	0.005	0.292
	Model B&C	0.782	0.131	0.052	0.021	0.008	0.003	0.003	0.363
t=10	Empirical Dist.	0.795	0.130	0.042	0.017	0.006	0.004	0.003	0.366
	Models B&C	0.820	0.086	0.045	0.023	0.012	0.006	0.008	0.377
t=20	Empirical Dist.	0.762	0.139	0.044	0.023	0.011	0.007	0.012	0.512
	Models B&C	0.808	0.080	0.047	0.027	0.016	0.009	0.013	0.458

and therefore underestimates the mean. Nevertheless, model B is a dramatic improvement over model A in terms of its accuracy at modeling the distribution of K. Figure 4.3 shows that the estimate of E[K] based on model B is much better than that based on model A, at all generations.

Table 4.5 shows the estimated transition matrices for generations 5, 10 and 20, for a population of constant size 100. Table 4.6 shows the empirical distributions of K based on 10,000 simulations, and the predicted distribution of K based on model B and the parameters of the estimated transition matrices shown in Table 4.5. The pattern is similar

Table 4.7: Mean and median lengths of different types of segments, based on 100,000 simulations, for a population of size $N = 20$.

	t=5		t=10		t=20	
Segment Type	mean	med.	mean	med.	mean	med.
U-U	0.105	0.073	0.055	0.038	0.030	0.021
T-U	0.108	0.075	0.056	0.039	0.031	0.021
U-V	0.105	0.075	0.057	0.040	0.031	0.021
T-V	0.098	0.069	0.053	0.036	0.032	0.022
V-T	0.126	0.084	0.070	0.047	0.042	0.028
V-S	0.150	0.102	0.084	0.058	0.051	0.035
S-T	0.155	0.104	0.086	0.058	0.051	0.036
S-S	0.167	0.110	0.093	0.062	0.052	0.035

to that observed for the smaller population. Model B overestimates the mean of K at generation 5, is quite close to the empirical value at generation 10, and underestimates the mean at generation 10. Figure 4.3 shows that model B is a great improvement over model A for N=100, also.

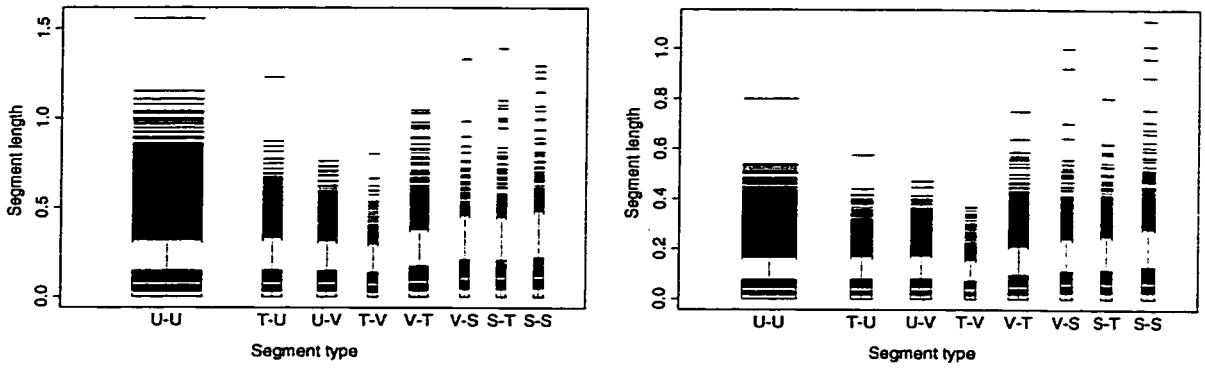
Despite this improvement in the component of the model that describes the sequence of junction types, the mean length of an IBD tract is still badly underestimated by model B. We therefore examine the other component of the model, the lengths of the segments. Under both model A and model B, the segment lengths are assumed to be the inter-arrival times of a Poisson process, and are therefore all identically exponentially distributed. Table 4.7 shows the mean and median lengths for the different segment types at generations 5, 10 and 20, as estimated from 100,000 simulations of a population of size N=20. The horizontal line across the table separates IBD segments from non-IBD segments. To get an idea of the variability in these estimates, compare segments of type U-V with segments of type T-U, and segments of type V-S with segments of type S-T. Within each pair, the mean and median lengths should be equivalent, since there is no inherent directionality in our conceptualization of the process. This comparison suggests that the estimates are adequate

Table 4.8: Mean and median lengths of different types of segments, based on 10,000 simulations, for a population of size $N = 100$.

	t=5		t=10		t=20	
Segment Type	mean	med.	mean	med.	mean	med.
U-U	0.101	0.069	0.051	0.036	0.027	0.019
T-U	0.113	0.075	0.056	0.038	0.027	0.019
U-V	0.098	0.062	0.051	0.035	0.027	0.019
T-V	0.157	0.112	0.060	0.041	0.023	0.018
V-T	0.107	0.060	0.055	0.035	0.030	0.019
V-S	0.148	0.104	0.076	0.055	0.039	0.027
S-T	0.139	0.077	0.090	0.063	0.035	0.024
S-S	0.128	0.100	0.079	0.051	0.045	0.030

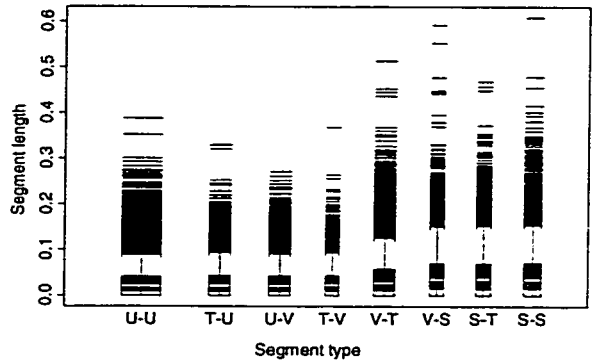
to the second decimal place. The estimates show that IBD segments are substantially longer than non-IBD segments. There also appears to be heterogeneity within IBD segments - segments of type S-S are longer on average than types V-S or S-T, which are in turn longer on average than segments of type V-T. To understand this, think of the process of crossovers along the chromosome. They are distributed according to a Poisson process with rate t per Morgan, where t is the generation number. Junctions are simply crossovers that happened in an area of non-IBD between the parent chromosomes. Junctions of type S are shared - the chromosomes are IBD at the location of that junction. Therefore, type S junctions tend to be in pieces of chromosome which are at higher frequency in the population. Therefore neighbouring crossovers are more likely to have occured in regions of IBD, and will not be visible as junctions, thus resulting in longer intervals.

Table 4.8 shows the mean and median length for each type of segment for a population of size $N = 100$, at generations 5, 10 and 20. These estimates are based on fewer simulations, because of the computational demands of simulating such a population, and there is therefore much greater variability in the estimates. The pattern of IBD segments being longer on average is still observable in generations 10 and 20.



(a) t=5

(b) t=10



(c) t=20

Figure 4.5: Distribution of segment length as a function of segment type, estimated by 100,000 simulations, for a population of size N=20 at generations (a) 5, (b) 10, and (c) 20.

Figure 4.5 shows boxplots of the simulation distribution of segment lengths according to segment type, for a population of size $N = 20$, at generations (a) 5 (b) 10 and (c) 20. The box shows the interquartile range of the simulated data, and the line denotes the median. The dotted lines extend out to 1.5 times the range, and data points outside that region are individually marked with a horizontal line. The width of each box is proportional to the square root of the number of data points it describes. The first four boxes correspond to non-IBD segments, and the last four boxes correspond to IBD segments. All segment types exhibit a form suggestive of the exponential distribution. IBD segments have slightly higher medians and heavier tails than non-IBD segments.

4.1.3 Model C: 1st order Markov dependence of junction types, IBD and non-IBD segments modelled separately

In model C, we incorporate some length differences into our model. We keep the first order Markov model for the sequence of junction types along the chromosome, but allow the IBD and non-IBD segment lengths to have different distributions. Assume that IBD segments are independent and identically distributed, with distribution F_I^s whose mean is μ_I. Also assume that non-IBD segments are independent and identically distributed, with distribution F_N^s whose mean is μ_N. Therefore model C describes the sequence of junction types along the chromosome as a Markov chain, with intervening segment lengths drawn from one of two distributions, depending on the type of the segment. This is an example of a semi-Markov process, which is a particular example of a renewal process. We apply some standard results of renewal theory to find equations for μ_I and μ_N which allow the estimation of $\mathrm{E}[L_I]$.

Consider now the lengths of IBD and non-IBD tracts under model C. The length of an IBD tract is the sum of a random number $K+1$ of segments, each of which has distribution F_I^s. The length of an IBD tract therefore has some distribution F_I which by Equation 4.20 has mean $(\mathrm{E}[K]+1)\cdot\mu_I$. Similarly, the length of a non-IBD tract has some distribution F_N, with mean $(\mathrm{E}[M]+1)\cdot\mu_N$. The sequence of IBD and non-IBD tracts along the chromosome is then an alternating renewal process. The limiting probability of IBD is then given by

the ratio of the mean length of an IBD tract to the sum of the means of IBD and non-IBD tracts(see for example Karlin & Taylor [18], p207), yielding the following equation:

$$1 - h_t(\underline{N}) = \frac{(\mathrm{E}[K]+1)\cdot\mu_I}{(\mathrm{E}[K]+1)\cdot\mu_I + (\mathrm{E}[M]+1)\cdot\mu_N} \tag{4.21}$$

Now, consider the sequence of V junctions along the chromosome. Their positions occur as a renewal process along the chromosome, with the intervening lengths having distribution $F_I + F_N$ with mean $(\mathrm{E}[K]+1)\cdot\mu_I + (\mathrm{E}[M]+1)\cdot\mu_N$. The same is true for type T junctions. Then for the stationary process, we have (Karlin & Taylor [18], p199)

$$\mathrm{E}[V_t] = \frac{1}{(\mathrm{E}[K]+1)\cdot\mu_I + (\mathrm{E}[M]+1)\cdot\mu_N} \tag{4.22}$$

and

$$\mathrm{E}[T_t] = \frac{1}{(\mathrm{E}[K]+1)\cdot\mu_I + (\mathrm{E}[M]+1)\cdot\mu_N}, \tag{4.23}$$

and so

$$\mathrm{E}[V_t + T_t] = \frac{2}{(\mathrm{E}[K]+1)\cdot\mu_I + (\mathrm{E}[M]+1)\cdot\mu_N}\ . \tag{4.24}$$

Equations 4.21 and 4.24 are easily solved to give the following equation for the length of an IBD tract:

$$\mathrm{E}[L_I] = (\mathrm{E}[K]+1)\cdot\mu_I = (1 - h_t(\underline{N}))\frac{2}{\mathrm{E}[V_t + T_t]} \tag{4.25}$$

We described Stam's method for calculating $\mathrm{E}[V_t + T_t]$ in the previous section, and recall that for a random mating population,

$$h_j(\underline{N}) = \prod_{i=0}^{j-1}\left(1 - \frac{1}{2N_i}\right)\ . \tag{4.26}$$

Thus the expected length of an IBD tract can be calculated, conditional on the population sizes.

Assessing model fit

Figure 4.2 shows the average length of an IBD tract (based on simulation) and the expected length of an IBD tract based on model C, for populations of size (a) N=20 and (b) N=100. Model C performs much better than both model A and model B. In the smaller population, model C still underestimates the mean somewhat after many generations. In the larger

population, there is initially a slight overestimate, but the estimate seems to be very good in later generations. This reflects the overestimation by the Markov model of $\mathrm{E}[K]$ in earlier generations, and underestimation in the later generations (see Figure 4.3).

Relationship to work of Stam [27]

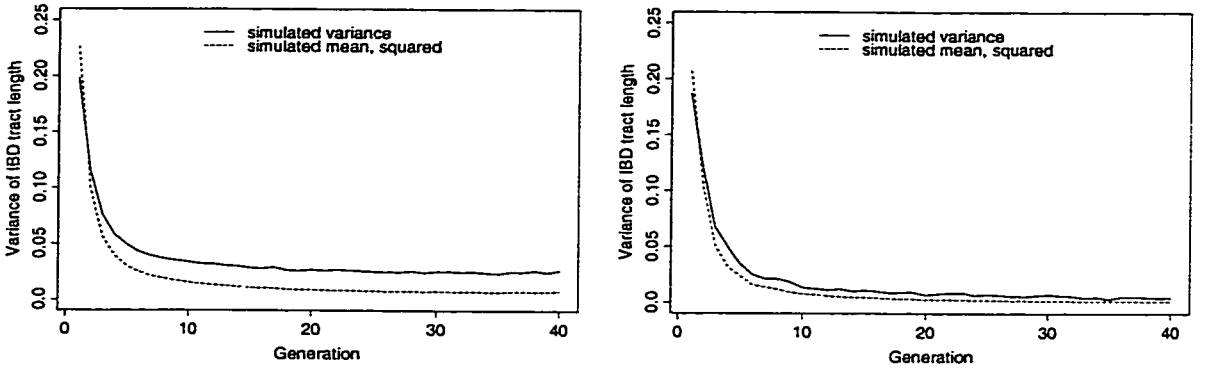
The goal of the Stam [27] paper in which he developed an expression for $\mathrm{E}[V_t + T_t]$ was to investigate the distribution of the fraction of the genome IBD in a finite random mating population such as the one which we model here. To this end, he assumed that the lengths of both IBD tracts and non-IBD tracts were exponentially distributed, with two different means. This is in contrast to our assumptions of model C, where IBD tracts have some distribution F_I, with mean $(\mathrm{E}[K]+1)\cdot\mu_I$, and non-IBD tracts have some distribution F_N, with mean $(\mathrm{E}[M]+1)\cdot\mu_N$. Stam's model gives rise to exactly the same estimate of the *mean* length of an IBD tract as our model C. However, Stam's exponential assumption implies that the variance of the length of an IBD tract must be the square of the mean.

Figure 4.6 shows the variance of the length of a random IBD tract, as estimated by simulations of a population of constant size (a) $N = 20$ and (b) $N = 100$. The dashed line shows the square of the mean, as estimated by simulation. Under Stam's exponential model, this should correspond to the variance. For the smaller population, Stam's exponential model seriously underestimates the variance. It performs better for the larger populations, but is still an underestimate. In view of this result, we explore the variance of the length of an IBD tract under our model C.

Modelling the variance of the length of an IBD tract

Recall that under the Markov model for the junction types that was used in models B and C, K (the number of type S junctions in a tract of IBD) has distribution given by Equation 4.6. We showed previously that

$$\mathrm{E}[K] = \frac{1-p_2}{\theta_2}. \tag{4.27}$$



(a) N=20

(b) N=100

Figure 4.6: Variance of the length of an IBD tract, as estimated directly by simulation, or as the square of the simulated mean, for populations of constant size (a) $N = 20$ and (b) $N = 100$.

Now

$$\mathrm{E}[K^2] = \sum_{k=1}^{\infty} k^2 \cdot (1-p_2)(1-\theta_2)^{k-1}\theta_2 = (1-p_2)\sum_{k=1}^{\infty} k^2(1-\theta_2)^{k-1}\theta_2 = \frac{(1-p_2)(2-\theta_2)}{\theta_2^2}. \tag{4.28}$$

Therefore the variance of K is given by

$$Var[K] = \frac{(1-p_2)(2-\theta_2)}{\theta_2^2} - \frac{(1-p_2)^2}{\theta_2^2}\ . \tag{4.29}$$

Let L_I denote the length of a randomly selected tract of IBD. Then $L_I = \sum_{i=1}^{K+1} X_i$, where K has mean and variance given in Equations 4.27 and 4.29 above, and under model C the X_i are independent identically distributed random variables with mean μ_I and variance σ_I^2. Then by conditional expectation,

$$\begin{aligned} Var[L_I] &= \mathrm{E}\left[Var[L_I|K=k]\right] + Var\left[\mathrm{E}[L_I|K=k]\right] \\ &= \mathrm{E}\left[(k+1)\cdot\sigma_I^2\right] + Var\left[(k+1)\cdot\mu_I\right] \\ &= \sigma_I^2\left(\mathrm{E}[K]+1\right) + \mu_I^2 Var[K+1] \\ &= \sigma_I^2\left\{\frac{1-p_2}{\theta_2}+1\right\} + \mu_I^2\left\{\frac{(1-p_2)(2-\theta_2)}{\theta_2^2} - \frac{(1-p_2)^2}{\theta_2^2}\right\}\ . \end{aligned} \tag{4.30}$$

If we now assume that the X_i have an exponential distribution with mean μ_I, then $\sigma_I^2 = \mu_I^2$ and we have

$$\begin{aligned} Var[L_I] &= \mu_I^2 \left\{ \frac{1-p_2}{\theta_2} + 1 + \frac{(1-p_2)(2-\theta_2)}{\theta_2^2} - \frac{(1-p_2)^2)}{\theta_2^2} \right\} \\ &= \mu_I^2 \left\{ \left(\frac{1-p_2}{\theta_2} + 1 \right)^2 + \frac{2(1-p_2)}{\theta_2} \left[\frac{p_2}{\theta_2} - 1 \right] \right\}. \end{aligned} \tag{4.31}$$

The first term of the variance is equivalent to the square of the mean, which represents the variance if the IBD tract lengths were exponentially distributed, as Stam assumed. In fact, the tract lengths would be exponentially distributed if the segment lengths were exponentially distributed and $K+1$ had a geometric distribution. However, $K+1$ has a zero-modified geometric distribution, and this gives rise to the additional variance which we see in the second term. The second term is positive, and therefore an increase in variance over Stam's exponential assumption if p_2 is greater than θ_2. Recall that p_2 is the probability that the present junction is of type T, given that the previous one was of type V. Also, θ_2 is the probability that the present junction is type T, given that the previous one was of type S. One might think of p_2 and θ_2 as representing the probability of terminating a short tract of IBD, versus the probability of terminating a longer tract of IBD. In our simulated examples, it was always the case that p_2 was greater than θ, and therefore it seems that our model gives an increase in variance relative to Stam's model, which we knew to underestimate the true variance. Unfortunately, we do not know of a theoretical approach to estimating $\frac{p_2}{\theta_2}$ under model C, so we cannot quantify the increase in variance predicted by our model.

4.2 Application to growing populations without subdivision

To demonstrate the potential effects of different types of population growth on the mean length of IBD tracts, we consider an example. We consider a population which has grown to one hundred times its initial size in one hundred generations. This time depth reflects the approximate age of the modern Finnish and Japanese populations, for example. We consider starting sizes of $N_0 = 20$, $N_0 = 100$, and $N_0 = 500$, and both linear and exponential growth. If the population is growing at a constant exponential rate, a 100 fold increase over 100 generations corresponds to a growth rate of 4.72% per generation. Figure 4.7 shows

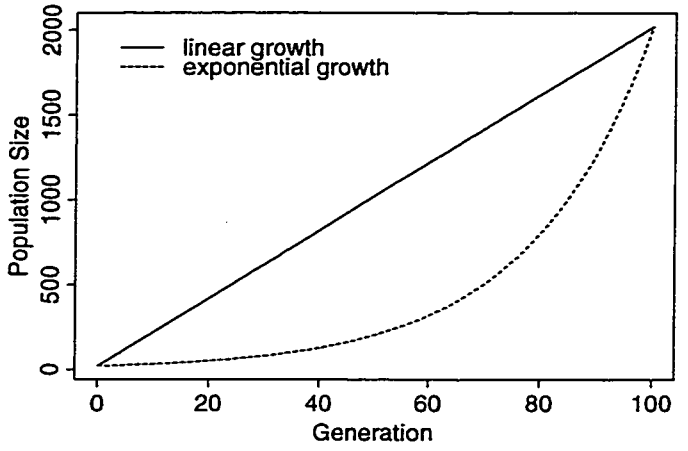
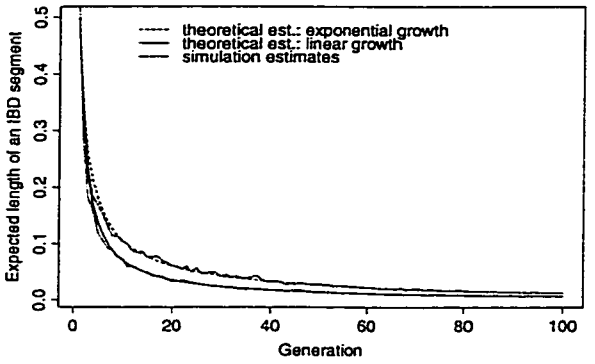


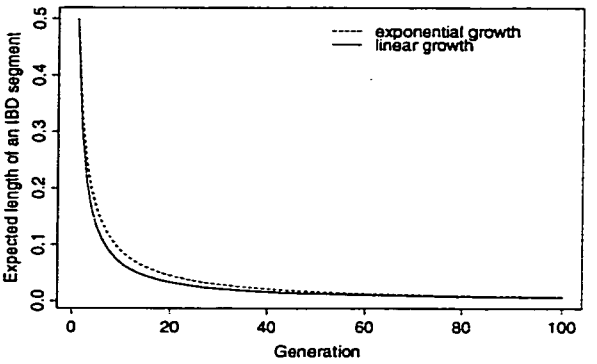
Figure 4.7: Population sizes over time for linear and exponentially growing populations.

population sizes over time for the exponentially and linearly growing populations. Neither scenario will reflect the exact history of any human population, but it is informative to consider these two idealized scenarios.

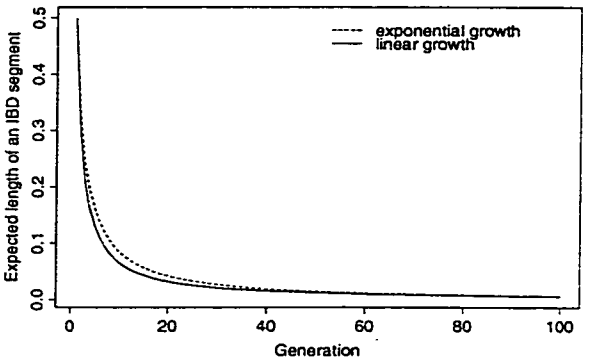
Figure 4.8 shows the expected length of an IBD tract based on model C, for exponentially and linearly growing populations with starting sizes of 20, 100 and 500. The plot for $N_0 = 20$ includes estimates of the mean length of an IBD tract in each of the two populations, based on 1000 simulations. The theoretical estimates agree well with the simulation based estimates, confirming that the approximations of model C work well in large populations such as these. We are particularly concerned with the length of an IBD tract in generation 100, since this mimics the situation where we examine a modern population whose age and founding size are approximately known, but whose exact growth pattern is not. Table 4.9 shows the expected length of an IBD tract in generation 100, in each of the populations considered. For the smaller population ($N_0 = 20$), there is a pronounced difference in the expected length of an IBD tract, depending on which type of growth the population has experienced. The expected length of an IBD tract in the exponentially growing population is 80% longer than in the linearly growing population. The increase is only 16% in the population with $N_0 = 100$, and there is virtually no difference in the largest population



(a) $N_0 = 20$



(b) $N_0 = 100$



(c) $N_0 = 500$

Figure 4.8: Expected length of an IBD tract, for populations expanding 100-fold over 100 generations, either linearly or exponentially, with (a) $N_0 = 20$, (b) $N_0 = 100$, and (c) $N_0 = 500$.

Table 4.9: Expected length of an IBD tract in two randomly chosen chromosomes from generation 100.

	N_0		
Growth type	20	100	500
Exponential	0.0129	0.0073	0.0065
Linear	0.0072	0.0063	0.0062

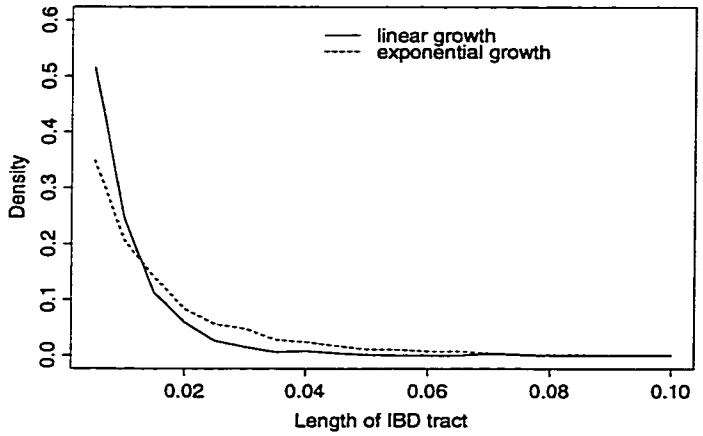


Figure 4.9: Estimated density, based on 1000 simulations, of the length of an IBD tract at generation 100, for exponentially and linearly growing populations with $N_0 = 20$ and $N_{100} = 2020$.

($N_0 = 500$).

In considering the importance of this increased length of IBD tracts in exponentially growing populations with small founder groups relative to linearly growing populations with small founder groups, it is useful to examine the full distribution of IBD tract lengths. Figure 4.9 shows the distribution of the length of an IBD segment at generation 100 for a population with sizes $N_0 = 20$ and $N_{100} = 2020$ (growing either exponentially or linearly) based on 1000 simulations. While there is clearly a difference between the means, the two distributions are very similar, and both have a high variance.

4.3 Application to growing populations with subdivision

4.3.1 Extension of Stam's work to subdivided populations

In order to consider the effect of population subdivision on the mean length of an IBD tract, we must extend Stam's method for calculating $\mathrm{E}[V_t + T_t|\underline{N}]$ so that it applies to subdivided populations. Consider a population that splits into two subpopulations at time t_s. Denote the population sizes over time in the two populations by $\underline{N}^1$ and $\underline{N}^2$. Now consider sampling one chromosome from each subpopulation at some time $t > t_s$. We need to calculate the expected number of type V and T junctions between these two chromosomes.

Let R_t^{112} denote the probability that three genes chosen randomly without replacement from generation t, two from population 1 and one from population 2, are all of different ancestral types. Then

$$R_{t+1}^{112} = \left(\frac{2N_t^1 - 1}{2N_t^1}\right) \cdot R_t^{112} = \left(1 - \frac{1}{2N_t^1}\right) \cdot R_t^{112} \tag{4.32}$$

since we require that the two genes from population 1 have distinct parents in generation t, and then that the three parents of the genes are themselves of different ancestral types. Similarly,

$$R_{t+1}^{122} = \left(1 - \frac{1}{2N_t^2}\right) \cdot R_t^{122} \,. \tag{4.33}$$

Now let Q_t^1 denote the number of V or T junctions existing on the chromosome from the first population, relative to the chromosome from the second population. Then $\mathrm{E}[V_t + T_t|\underline{N}] = \mathrm{E}[Q_t^1|\underline{N}^1] + \mathrm{E}[Q_t^2|\underline{N}^2]$, and analogous to Equation 4.18, we have

$$\mathrm{E}[Q_{t+1}^1|\underline{N}] = \mathrm{E}[Q_t^1|\underline{N}] + h_t^1(\underline{N}^1) - R_t^{112}, \tag{4.34}$$

where $h_t^1(\underline{N}^1)$ denotes the probability of non-IBD within subpopulation 1. The first term in the equation corresponds to junctions that already existed in the previous generation. The second term adds in the expected number of newly formed junctions in population 1, and the third term subtracts out the expected number of those which are of type U. analogously,

$$\mathrm{E}[Q_{t+1}^2|\underline{N}] = \mathrm{E}[Q_t^2|\underline{N}] + h_t^2(\underline{N}^2) - R_t^{122}. \tag{4.35}$$

These equations can be applied iteratively to calculate $\mathrm{E}[V_t + T_t|\underline{N}]$.

Recall Equation 4.25 for the expected length of an IBD tract:

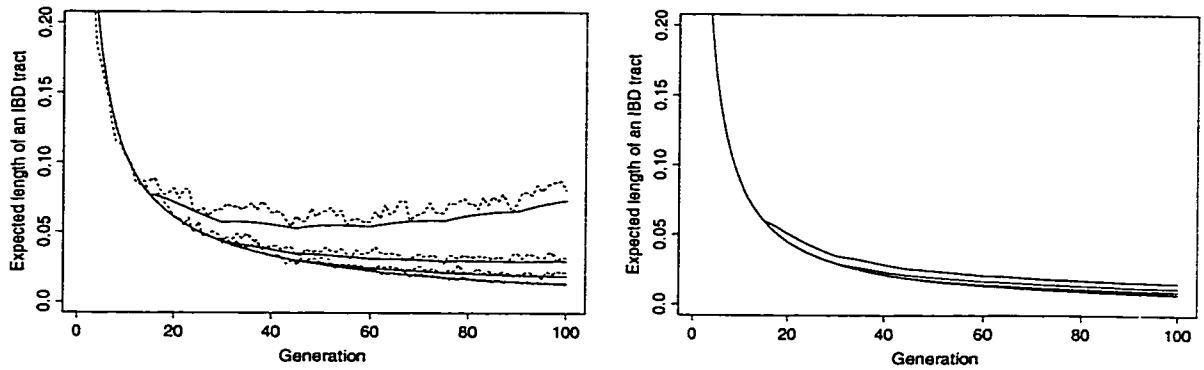
$$\mathrm{E}[L_I] = (1 - h_t(\underline{N}))\frac{2}{\mathrm{E}[V_t + T_t]} \tag{4.36}$$

When one is comparing chromosomes from each of two subpopulations, $h_t(\underline{N})$ denotes the probability of IBD for two genes sampled on each from the subpopulations. If we denote this quantity by $h_t^{12}(\underline{N})$ then $h_t^{12}(\underline{N}) = h_{t_s}(\underline{N})$. To understand this relationship, think of the genes sampled from the two populations at time t. Since there is no migration between the populations, these genes are copies of two distinct genes present in the population at time t_s. Using these extensions to Stam's theory, we can calculate the expected length of an IBD tract between two chromosomes from different subpopulations.

4.3.2 Example of the effects of population subdivision

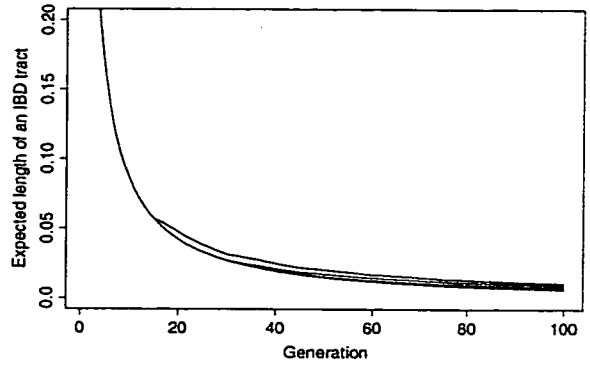
To demonstrate the potential effects of differing degrees of population subdivision on the mean length of IBD tracts, we consider an example. We consider a population which has grown to one hundred times its initial size in one hundred generations, by constant exponential growth. We consider starting sizes of $N_0 = 20$, $N_0 = 100$, and $N_0 = 500$, and subdivision scenarios in which populations repeatedly bifurcate when the local population size reaches $2N_0$, $4N_0$ or $8N_0$. These scenarios correspond to a total population which is divided into 64, 32, or 16 subpopulations at generation 100, respectively. For all scenarios with a given starting population size, the total population size is the same at all generations - the only difference is in the degree of subdivision within the total population.

Figure 4.10 shows the expected length of an IBD tract between two chromosomes from within the same subpopulation. Different panels correspond to the three different starting population sizes. All populations are growing at a constant exponential rate of 4.72% per generation. Within each panel, the lines from top to bottom correspond to the most subdivided population to the least subdivided population. The bottom line is the population with the same total size, but with no subdivision. Estimated mean lengths based on 1000 simulations are also shown for the population with starting size $N_0 = 20$. Larger populations were not simulated because of computational demands, and the results of the previous sections suggest that approximation will be better for larger populations. Expected IBD tract



(a) $N_0 = 20$

(b) $N_0 = 100$

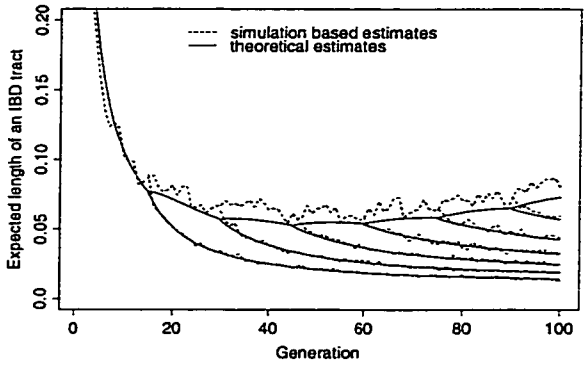


(c) $N_0 = 500$

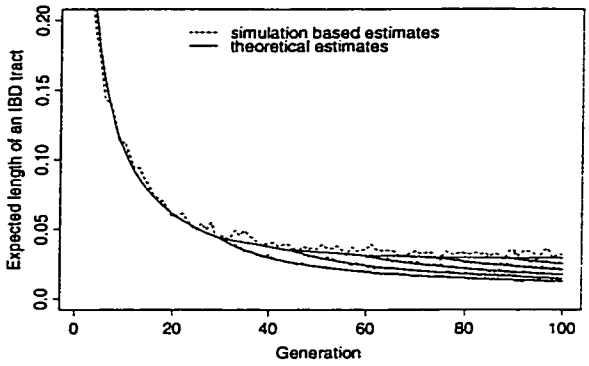
Figure 4.10: Expected length of an IBD tract, between two chromosomes from within the same sub-population. Different lines within a plot correspond to populations with different amounts of subdivision. The lines from the top go from most subdivided to least subdivided, with the bottom line being the non-subdivided population. The three plots correspond to different starting sizes of (a) 20, (b) 100, and (c) 500. All populations are growing at a constant exponential rate of 4.72% per generation.

lengths within and between sub-populations at generation 100 are tabulated in Tables 4.10, 4.11 and 4.12. For the smallest population, the level of subdivision has a pronounced effect on the expected length of an IBD tract. We consider in particular the expected length of an IBD tract at generation 100. For the most subdivided case, where the first subdivision occurs at $t = 15$, the expected length of an IBD tract is 0.0729, relative to 0.0129 in the undivided case. This is an extreme level of subdivision, but even for the least subdivided case, where the first subdivision occurs at $t = 45$, the expected length is 0.0184, 50% larger than in the undivided population. The potential effects of subdivision are greatly reduced in the two larger populations. In the population with $N_0 = 100$, the most subdivided population has mean IBD length twice as large as the undivided population. When $N_0 = 500$, the expected length of an IBD tract in the most subdivided population is 65% higher than that in the undivided population.

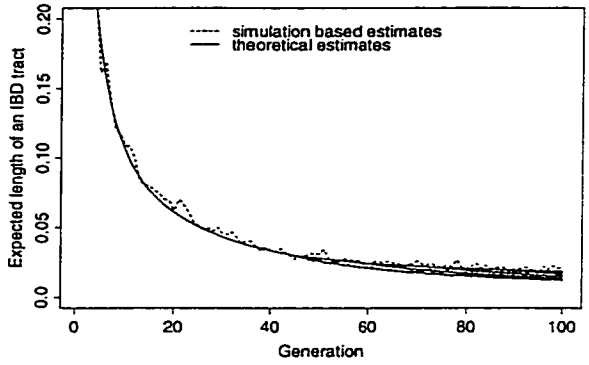
Figure 4.11 shows the expected length of an IBD tract, for the population with starting size $N_0 = 20$. Different panels in the figure correspond to different levels of subdivision. The smaller the splitting size of the population, the more subdivisions there are at any time t, and in particular at $t = 100$. Within each plot, the top line shows the expected length of an IBD tract within a subdivision. The other lines show the expected length of an IBD tract between two subdivisions. For example, in panel (a), the first line to diverge from the top line does so at $t = 15$, and shows the expected length of an IBD tract between two chromosomes from subpopulations that split from one another at generation $t = 15$. The numbers for generation $t = 100$ over all three panels are summarized in Table 4.10. For the most subdivided population (i.e. split size of 40), there is a substantial difference in expected IBD tract length between chromosomes from different subpopulations, depending on when the populations split from one another. For example, at generation 100, the expected length of an IBD tract for a pair of chromosomes whose populations split at generation 15 is 0.0144, whereas if the populations split at generation 90, it would be 0.0570. IBD tracts in chromosomes whose populations split at generation 90 are on average 4 times the length of IBD tracts in chromosomes whose populations split at generation 15. Differences become less pronounced, although they still exist, as the populations become less subdivided (split sizes of 80 and 160).



(a) split size 40



(b) split size 80



(c) split size 160

Figure 4.11: Expected length of an IBD tract, for an exponentially growing population with $N_0 = 20$ and $N_{100} = 2020$, and different levels of subdivision. The populations bifurcate whenever (a) $N = 40$, (b) $N = 80$, or (c) $N = 160$.

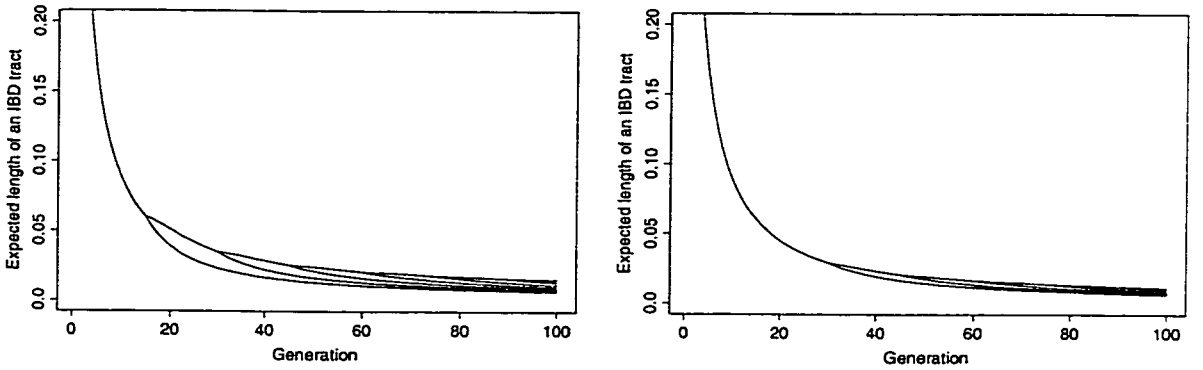
Table 4.10: Expected length of an IBD tract for chromosomes chosen within or between sub-populations at generation 100, for an exponentially growing population with $N_0 = 20$ and $N_{100} = 2020$.

	Between sub-pops which split at:						Within
	$t = 15$	$t = 30$	$t = 45$	$t = 60$	$t = 75$	$t = 90$	sub-pop
split size 40	.0144	.0189	.0248	.0324	.0426	.0570	.0729
split size 80	-	.0126	.0147	.0174	.0208	.0251	.0294
split size 160	-	-	.0122	.0134	.0149	.0167	.0184
no subdivision	-	-	-	-	-	-	.0129

Table 4.11: Expected length of an IBD tract for chromosomes chosen within or between sub-populations at generation 100, for an exponentially growing population with $N_0 = 100$ and $N_{100} = 10,100$.

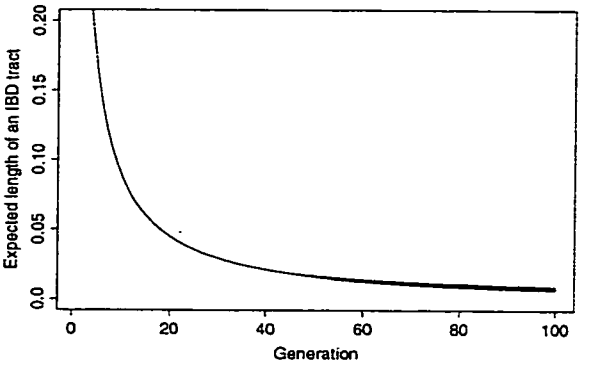
	Between sub-pops which split at:						Within
	$t = 15$	$t = 30$	$t = 45$	$t = 60$	$t = 75$	$t = 90$	sub-pop
split size 200	.0066	.0075	.0085	.0098	.0113	.0132	.0150
split size 400	-	.0067	.0073	.0081	.0091	.0103	.0114
split size 800	-	-	.0068	.0073	.0078	.0085	.0092
no subdivision	-	-	-	-	-	-	.0073

Figure 4.12 shows the expected length of an IBD tract, for the population with starting size $N_0 = 100$. Notation is the same as Figure 4.11. The numbers for generation $t = 100$ over all three panels are summarized in Table 4.11. The same patterns that we observed for the smallest populations (i.e. those with $N_0 = 20$) are visible here. In particular, we see longer tracts of IBD sharing between more recently diverged populations than we do between populations which diverged many generations ago. These differences are successively less pronounced in populations with less subdivision (split sizes 400 and 800). Even for the most subdivided population, the differences are much smaller than those observed in the populations with the smallest founding size.



(a) split size 200

(b) split size 400



(c) split size 800

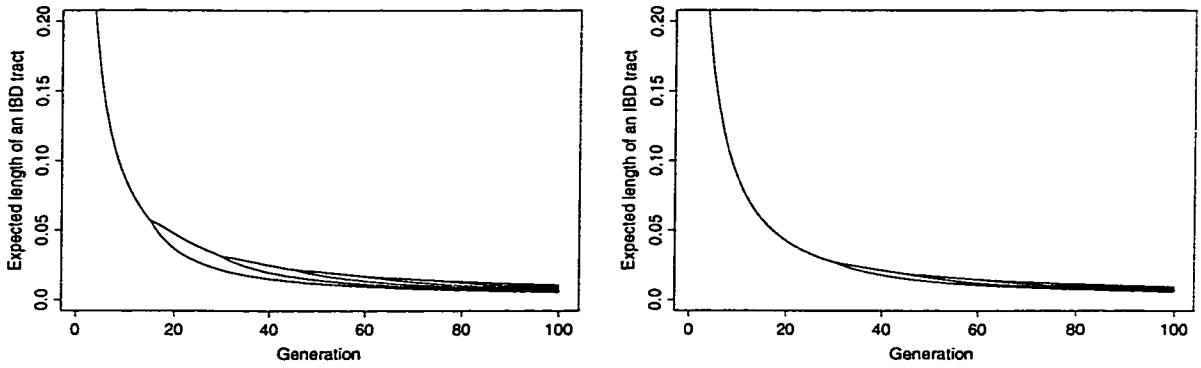
Figure 4.12: Expected length of an IBD tract, for an exponentially growing population with $N_0 = 100$ and $N_{100} = 10,100$, and different levels of subdivision. The populations bifurcate whenever (a) $N = 200$, (b) $N = 400$, or (c) $N = 800$.

Table 4.12: Expected length of an IBD tract for chromosomes chosen within or between sub-populations at generation 100, for an exponentially growing population with $N_0 = 500$ and $N_{100} = 50,500$.

	Between sub-pops which split at:						Within
	$t = 15$	$t = 30$	$t = 45$	$t = 60$	$t = 75$	$t = 90$	sub-pop
split size 1000	.0056	.0061	.0067	.0075	.0084	.0096	.0107
split size 2000	-	.0058	.0063	.0069	.0076	.0085	.0093
split size 4000	-	-	.0060	.0064	.0069	.0074	.0079
no subdivision	-	-	-	-	-	-	.0065

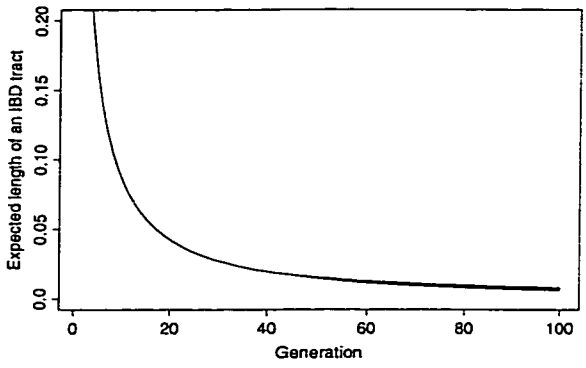
Figure 4.13 shows the expected length of an IBD tract, for the population with starting size $N_0 = 500$. Notation is the same as Figures 4.11 and 4.12. The numbers for generation $t = 100$ over all three panels are summarized in Table 4.12. These largest populations ($N_0 = 500$) confirm the trend observed in the smaller populations: the larger the population, the less effect subdivision has on the expected lengths of within and between population IBD sharing.

The above results show that population subdivision can have a pronounced effect on the mean lengths of IBD tracts within and between subpopulations, particularly when overall population sizes are small and there is a lot of subdivision. It must be kept in mind, however, that these observations are based on theoretical means, and the distributions themselves are quite variable. Figure 4.14 shows the estimated distribution, based on 1000 simulations, of the length of a random IBD tract between two chromosomes chosen from within the same subdivision, at generation 100, of populations with same total size, but varying levels of subdivision. These populations all have a starting size of $N_0 = 20$, since this is the example in which the differences are the strongest (see Table 4.10). While the differences in mean can clearly be seen, all four distributions have a large variance, and their ranges overlap. This implies that while the means are quite different, it would be hard to distinguish between the different types of populations without large amounts of data.



(a) split size 1000

(b) split size 2000



(c) split size 4000

Figure 4.13: Expected length of an IBD tract, for an exponentially growing population with $N_0 = 500$ and $N_{500} = 50,500$, and different levels of subdivision. The populations bifurcate whenever (a) $N = 1000$, (b) $N = 2000$, or (c) $N = 4000$.

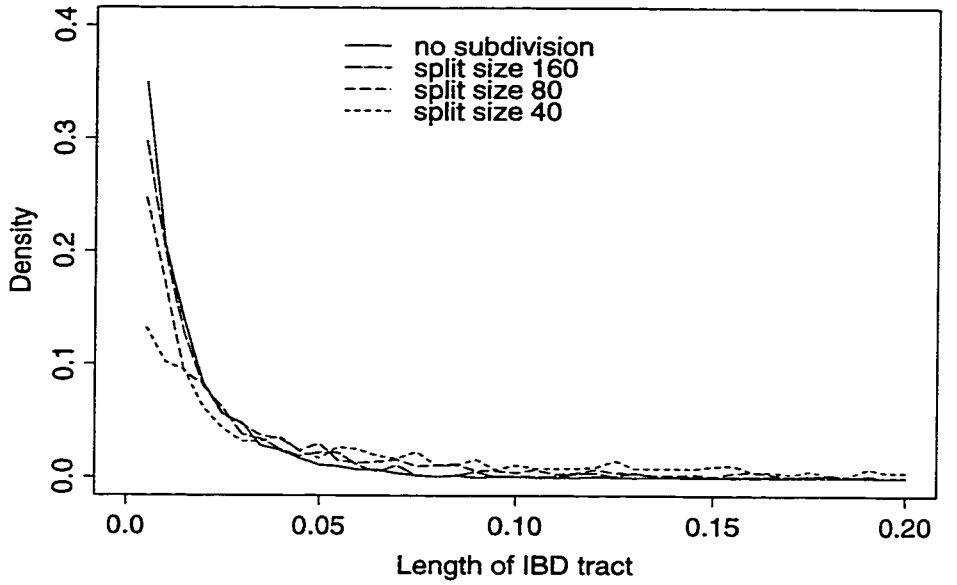


Figure 4.14: Estimated density, based on 1000 simulations, of the length of an IBD tract at generation 100, for exponentially growing populations with $N_0 = 20$, $N_{100} = 2020$, and varying levels of subdivision.

4.4 Discussion

In this chapter, we modelled the length of a tract IBD between two chromosomes. The best fitting model is one that allows for 1st order Markov dependence of junction types along the chromosomes, and allows different length distributions for IBD and non-IBD segments. We calculated the mean length of an IBD tract based on this model, and showed that the variance of the length of an IBD tract is greater than the square of the mean. Using these results, we investigated the effects of different types of growth on the expected length of an IBD tract. Similar to the results of previous chapters, we found that different types of population growth do affect the mean length of an IBD segment, but the effect is large only in small populations. Our extension of Stam's calculation of the expected number of external junctions between two chromosomes to subdivided populations allowed us to consider the effects of population subdivision on the expected length of an IBD tract both within and between subpopulations. We found that subdivision does affect the mean length of an IBD tract, both within and between subpopulations, but the effect is large only in small populations.

Chapter 5

SIMULATING CHROMOSOME TRANSMISSIONS IN RANDOM MATING POPULATIONS AND ON THE HUTTERITE ANCESTRY

In this chapter, we discuss some of the computational issues associated with simulating chromosome transmissions in random mating populations or in large pedigrees. We first outline the important data structures and algorithms used in our programs. The second part of the chapter presents some results of simulations on the Hutterite ancestry. In particular, we examine the number of junctions in a randomly selected chromosome from the most recent generation of Hutterites, and the lengths of IBD tracts between two chromosomes chosen from (i) the same colony; (ii) different colonies within the same leut; and (iii) different leut. In each case, the simulation results are compared to theoretical predictions made using the results of chapters 2, 3, and 4, assuming a random mating population with the same historical sizes and history of subdivision as the Hutterites. To assess the importance of any observed differences, we also simulate a random mating population with the same historical sizes and history of subdivision as the Hutterites.

5.1 Data Structures

Simulation programs were written in C, in part because of the flexibility this language offers in terms of defining data structures. In particular, we defined structures to represent individuals, and the segments of different ancestral types which make up a chromosome.

5.1.1 Individuals

An individual is represented by a structure containing an identifying number, the identifying numbers of the individual's parents, and two pointers - one to the individual's maternal chromosome, and one to their paternal chromosome. The structure may also contain mis-

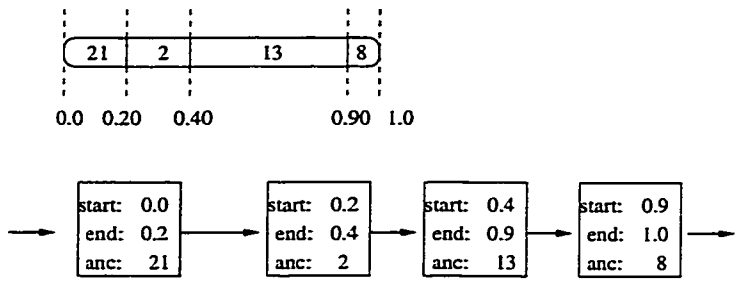


Figure 5.1: Schematic of chromosome representation.

cellaneous information about the individual - examples include sex, date of birth, and leut. If one wished to model the transmission of two unlinked chromosomes, one would simply add two more pointers to the structure - one each to the maternal and paternal copies of the second chromosome.

5.1.2 Segments and Chromosomes

Chromosomes are represented by a linked list of segments. A segment is a structure made up of two real numbers, representing the beginning and end of the segment (in Morgans) and an integer reflecting its ancestral origin in the founder population. Figure 5.1 shows an example of how a chromosome would be represented. The chromosome shown has total length one Morgan, and is made up of four segments from ancestral chromosome types 21, 2, 13 and 8. The junctions between these types occur at 0.2M, 0.4M and 0.9M. This is then represented as a linked list of four segments, as shown in the figure. For a given individual, the pointers to his or her maternal and paternal chromosomes are therefore pointers to the first segment in such a list.

5.2 Algorithms

In this section, we discuss the algorithms used for gamete production, simulation of a random mating population, and simulations on a pedigree structure. We also give examples of the computational demands of our programs.

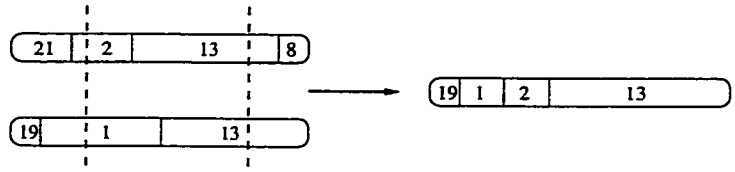


Figure 5.2: Schematic of gamete production.

5.2.1 Gamete Production

In order to simulate the production of a gamete from an individual, the locations of recombinations along the chromosome are generated, according to a Poisson process with rate one per Morgan. One can think of copying information from the parent chromosomes into a gamete chromosome. First, one of the two parent chromosomes is randomly selected. Information is copied from this chromosome until the first recombination event is encountered, at which point we switch to copying information from the other parent chromosome. We continue, switching at recombination events, until the end of the chromosome is reached. This representation is due to Robertson [26]. Figure 5.2 shows the production of a gamete from two parent chromosomes. The vertical dashed lines represent the locations where recombinations occur, as simulated from the Poisson process. The bottom parent chromosome was randomly selected to begin with, and the chromosome was copied up until the first recombination event, where we switched to the top parent chromosome. This chromosome was then copied until the next recombination event, at which position we switched back to the bottom parent chromosome. Note that one new junction was created in this meiosis, at the first recombination event between ancestral chromosome types 1 and 2. A junction was not formed at the second recombination event, since the two parent chromosomes were IBD at that point. The existing junctions between types 19 and 1 and types 2 and 13 were passed on, while the junctions between type 21 and type 2, and type 13 and 8 were not.

5.2.2 Random Mating Populations

In our simulations of random mating populations, we assume that population sizes are known at every generation, and that the members of the founding population are all non-

inbred and unrelated. Founders therefore have intact chromosomes of distinct ancestral types. For each individual in generation $j+1$, two parents are randomly chosen, with replacement, from the individuals in generation j. A gamete is generated from each of these parents, and the gametes together make up the chromosomal complement of the offspring. This is repeated until the desired population size has been achieved for generation $j+1$. The whole process is repeated for each generation, for as many generations as is required.

To minimize memory use, only one generation of individuals are stored at any one time. As soon as generation $j+1$ is complete, the individuals in generation j are deleted. The individuals in generation $j+1$ are then analysed as required, and we move on to the simulation of generation $j+2$.

5.2.3 Pedigree Based Simulation

A pedigree is represented as an array of individuals. Founding individuals are assumed to be non-inbred and unrelated, and therefore have intact chromosomes of distinct ancestral types. Individuals are stored in the pedigree such that parents always precede children. This allows simulation to proceed by stepping through the individuals in the pedigree, simulating chromosome data for each individual in turn, conditional on already simulated chromosome data in their parents.

5.2.4 Computational Demands

Simulating random mating populations over long periods of time can be quite computationally demanding. For example, consider a chromosome in generation 100 of a large population. Such a chromosome will on average be made up of about 101 segments per Morgan. Therefore each chromosome is a list about $101L$ segments long, where L is the length of the chromosome. All the simulations described in this chapter used chromosomes of length one Morgan. Every individual has two chromosomes, and so in a large population, memory demands become quite large. Table 5.1 shows some examples of run-time and memory usage on a Pentium II 333MHz processor with 96MB of RAM. Simulations on the Hutterite pedigree were run on a different machine, and run-times and memory usage will

Table 5.1: Examples of computational demands of simulation of random mating populations.

Population Description	Number of Generations	Number of Realizations	Time	Memory Use
$N = 20$, constant size	40	100,000	3h 10min	0.2 MB
$N = 100$, constant size	40	10,000	2h 23min	1.1 MB
$N_0 = 20$, $N_{100} = 2020$, exponential growth	100	1,000	8h 6min	46 MB
$N_0 = 20$, $N_{100} = 2020$, linear growth	100	1,000	25h 49min	57 MB

be reported in the following sections.

5.3 Simulations on the Hutterite Ancestry

5.3.1 Hutterite History, Condensed

The Hutterites are one of several anabaptist groups that arose during the Protestant Reformation of the 16th century. The first Hutterite colony was founded in 1528 in Moravia, and by 1593 there were approximately 20,000 Hutterites, living in about 80 colonies [16]. In 1593 war broke out between Austria and Turkey, and raids by both sides destroyed many Hutterite colonies. The Thirty Years War followed and by 1622, few Hutterites remained.

In 1770 a group of 123 Hutterite refugees settled in the area of Russia which is now the Ukraine [16]. They were there by invitation of a Russian general, and had been promised freedom to practise their religion as they wished. By 1842, the Hutterite population consisted of 384 people (185 men and 199 women) in 69 families [15]. The population continued to grow, and by 1868, there were five Hutterite villages. Internal strife had caused the Hutterites to give up communal living in 1819. In 1859 the commitment to communal living was renewed in part of one of the villages. This group became known as the Schmiedeleut. Another group from the same village followed suit the following year, and became known as the Dariusleut. Shortly thereafter in 1872 the Russian government repealed laws that

had granted religious groups including the Hutterites exemption from military training, an act which prompted the departure of all Hutterites from Russia. The Schmiedeleut and the Dariusleut left Russia in 1874, and founded the Bonne Homme and Silver Lake colonies respectively, both in South Dakota. A third group left Russia in 1877 and founded the Elmspring colony in South Dakota. This group became known as the Lehrerleut. A total of 1265 Hutterites had come to North America by 1879 [15]. Approximately two-thirds of this group settled on single-family farms, and the rest made up the three founding colonies from which the modern Hutterite population is descended. The Dariusleut moved to Canada to avoid participation in the Spanish-American war in 1899. The outbreak of World War I increased hostility towards Germans in the USA, and prompted a mass migration of Hutterites to Canada in 1918. By World War II, when there was a more tolerant attitude towards conscientious objectors, there were a total of 56 Hutterite colonies in Canada, and 6 in the United States [16].

5.3.2 The Hutterite Pedigree

Through collaboration with Dr. Ken Morgan, of the University of Montreal, we have access to a pedigree which traces the ancestry of the cohort of Hutterites alive in 1981 back to ancestors who were born in the 1700s. There are a total of 27,163 individuals in the pedigree, of whom 130 are founders. Seventy one of those founders were born prior to or during 1800. Information available on each of the individuals includes year of birth, sex, and vital status as of December 31st, 1981. We also have leut and coded colony information for most individuals, which is defined as either leut/colony of birth, or for those not born in a colony, leut/colony of residence in 1880.

We simulated the transmission of a chromosome of length one Morgan down this pedigree, assuming that all founder individuals were non-inbred and unrelated. Three colonies were chosen from each leut. Colonies chosen were required to have a substantial number of members alive in the 1981 census, and an attempt was made to select colonies with ordered relationships between them. For example, of the selected Lehrerleut colonies (L1, L2 and L3), L2 and L3 divided in about 1940, and the ancestral colony of L2 and L3 divided from

colony L1 in about 1900. The same general relationships are true of selected colonies S1-S3 in the Schmiedeleut, and D1-D3 in the Dariusleut. Within each of these nine colonies, three individuals were randomly selected. For each of the 27 individuals, the number of junctions in each of their two chromosomes was recorded, for each iteration. To investigate the patterns of IBD sharing between individuals, the paternal chromosome of each of the 27 individuals was compared in turn to the paternal chromosome of all 26 other individuals. If there existed one or more tracts of IBD between the two chromosomes, a tract was randomly selected, and its length recorded for each iteration. The simulation consisted of 100,000 iterations, which took 21 hours and 16 minutes to run, using about 22.4MB of memory, on a dual 667 MHz processor Alpha with 2GB of memory.

5.3.3 Approximating the Hutterite population by a random mating population

The theoretical results presented in Chapters 2, 3, and 4 are for discrete generation random mating populations in which the population sizes over time can be specified. In order to apply our theory to the Hutterite population, we must approximate it by such a population. To estimate historical population sizes, we classified all individuals in the ancestry into 20 year birth cohorts, starting in 1980 and working backwards. Table 5.2 shows these population sizes, and the number of founders and members of each of the three leut within each. We decided to use the 1740 to 1760 birth cohort as the "founding" generation. Since the migration to the Ukraine was complete by 1770, this cohort can be thought of as the parents of those settlers. Furthermore, we used the 1840 to 1860 birth cohort as the generation in which the subdivision into three leut occured. This cohort can be thought of as the parents of those born between 1860 and 1880, who would likely have been the group which re-established communal living prior to the move to North America.

In order to use these cohort sizes as the sizes of a subdivided, random mating, discrete generation population, we must assign leut to those for whom it is missing and who were born after 1840. If leut information was missing for a parent who was born after 1840, they were assigned the leut of their children. This rule alone resolved all but 26 of the individuals for whom leut information was missing. Of these 26, 11 were founders with no

Table 5.2: Number of individuals in 20 year birth cohorts, classified according to founder status and known leut information.

Birth Cohort	Total	Founders	Missing Leut	L-leut	S-leut	D-leut
$yob \leq 1700$	3	3	3	0	0	0
$1700 < yob \leq 1720$	16	16	16	0	0	0
$1720 < yob \leq 1740$	27	19	27	0	0	0
$1740 < yob \leq 1760$	55	22	55	0	0	0
$1760 < yob \leq 1780$	58	8	58	0	0	0
$1780 < yob \leq 1800$	64	3	64	0	0	0
$1800 < yob \leq 1820$	98	2	78	2	9	9
$1820 < yob \leq 1840$	117	8	69	17	17	14
$1840 < yob \leq 1860$	157	3	54	18	48	37
$1860 < yob \leq 1880$	276	3	57	65	90	64
$1880 < yob \leq 1900$	609	3	51	168	203	187
$1900 < yob \leq 1920$	1409	7	48	428	505	428
$1920 < yob \leq 1940$	3257	9	14	1058	1249	936
$1940 < yob \leq 1960$	7477	16	16	2163	3189	2109
$1960 < yob \leq 1980$	12805	8	12	3545	5557	3691
$1980 < yob \leq 2000$	735	0	0	206	322	207

Table 5.3: Population sizes used for a discrete generation, subdivided, randomly mating population approximating the Hutterites, based on 20 year birth cohorts.

Gen.	Birth cohort	Pop. Size	L-leut	S-leut	D-leut
0	$1740 < yob \leq 1760$	55	0	0	0
1	$1760 < yob \leq 1780$	58	0	0	0
2	$1780 < yob \leq 1800$	64	0	0	0
3	$1800 < yob \leq 1820$	98	0	0	0
4	$1820 < yob \leq 1840$	117	0	0	0
5	$1840 < yob \leq 1860$		28	77	52
6	$1860 < yob \leq 1880$		77	111	86
7	$1880 < yob \leq 1900$		182	227	200
8	$1900 < yob \leq 1920$		434	541	433
9	$1920 < yob \leq 1940$		1058	1257	940
10	$1940 < yob \leq 1960$		2164	3193	2115
11	$1960 < yob \leq 1980$		3545	5557	3691

children, and were therefore removed from the pedigree. Eleven others were members of 3 distinct nuclear families, who were apparently living with the Hutterites at the time of the census, but were unrelated to them. This group was also removed. Finally, the remaining 4 individuals were clearly Hutterites, but were childless and had no leut information. They were assigned the leut of the closest relative possible - in all cases this was either a niece or a nephew. Table 5.3 shows the populations sizes used for a discrete generation, subdivided, randomly mating population approximating the Hutterites, based on 20 year birth cohorts and leut assignments as described above. This model will be referred to in what follows as random-20.

To investigate the sensitivity of our theoretical calculations to different ways of approximating the Hutterite's population history, we also considered a random mating population whose population sizes over time were based on 25 year birth cohorts. Table 5.4 shows the resulting cohort sizes, and the number of founders and members of each of the three leut

Table 5.4: Number of individuals in 25 year birth cohorts, classified according to founder status and known leut information.

Birth Cohort	Total	Founders	Missing Leut	L-leut	S-leut	D-leut
$yob \leq 1705$	5	5	5	0	0	0
$1705 < yob \leq 1730$	29	27	29	0	0	0
$1730 < yob \leq 1755$	54	24	54	0	0	0
$1755 < yob \leq 1780$	71	12	71	0	0	0
$1780 < yob \leq 1805$	83	4	83	0	0	0
$1805 < yob \leq 1830$	140	7	102	9	16	13
$1830 < yob \leq 1855$	167	5	66	22	42	37
$1855 < yob \leq 1880$	322	3	71	71	106	74
$1880 < yob \leq 1905$	862	4	66	245	288	263
$1905 < yob \leq 1930$	2421	7	39	779	899	704
$1930 < yob \leq 1955$	7001	21	21	2050	2916	2014
$1955 < yob \leq 1980$	15273	11	15	4288	6600	4370
$1980 < yob \leq 2005$	735	0	0	206	322	207

Table 5.5: Population sizes used for a discrete generation, subdivided, randomly mating population approximating the Hutterites, based on 25 year birth cohorts.

Gen.	Birth cohort	Pop. Size	L-leut	S-leut	D-leut
0	$1730 < yob \leq 1755$	54	0	0	0
1	$1755 < yob \leq 1780$	71	0	0	0
2	$1780 < yob \leq 1805$	83	0	0	0
3	$1805 < yob \leq 1830$	140	0	0	0
4	$1830 < yob \leq 1855$		35	73	59
5	$1855 < yob \leq 1880$		86	132	102
6	$1880 < yob \leq 1905$		263	320	278
7	$1905 < yob \leq 1930$		781	932	708
8	$1930 < yob \leq 1955$		2051	2922	2023
9	$1955 < yob \leq 1980$		4288	6601	4370

within each. We decided to use the 1730 to 1755 birth cohort as the "founding" generation, and the 1830 to 1855 birth cohort as the generation in which the subdivision into three leut occurred, again attempting to match the approximate time of settlement in the Ukraine, and the approximate time of leut formation, respectively.

We used the same approach to assigning missing leut as previously described for the model based on 20 year birth cohorts, and thereby obtained the population sizes shown in Table 5.5. This model will be referred to as model random-25 in what follows. Using a larger generation time (25 years rather than 20) results in a population which is nine generations from founding, rather than eleven. Correspondingly, the population sizes within each generation are larger. We will discuss the implications of these differences in the following sections.

Table 5.6: Expected number (standard deviation) of junctions per Morgan in a chromosome randomly selected from the most recent generation of L-leut, S-leut and D-leut, based on models random-20 and random-25.

	L-leut	S-leut	D-leut
$E[J_9 \mid \underline{N}]$, random-25	8.75 (2.96)	8.78 (2.96)	8.77 (2.96)
$E[J_{11} \mid \underline{N}]$, random-20	10.59 (3.25)	10.65 (3.26)	10.63 (3.26)

5.3.4 Number of junctions in a randomly selected chromosome

Theoretical predictions based on Hutterite population sizes

Table 5.6 shows the expected number of junctions per Morgan in a randomly selected chromosome from the most recent generation of Hutterites, based on models random-20 and random-25. Also shown is an estimate of the standard deviation of the number of junctions, based on the variance approximation developed in Chapter 2. The approximation should be quite accurate, since the populations considered here are young and rapidly growing.

Two things are apparent from Table 5.6; (i) within either model, the expected numbers of junctions in chromosomes from each of the three leut are very similar; and (ii) the two models give quite different numbers of expected junctions (about 8.8 vs. 10.6). To understand these observations, we consider the calculation of the expected number of junctions. For a random mating population without subdivision,

$$E[J_t \mid \underline{N}] = \sum_{j=0}^{t-1} h_j(\underline{N}) L \tag{5.1}$$

(Equation 2.6), where $h_j(\underline{N})$ is the probability of non-IBD at a particular locus for an individual in generation j, where $\underline{N}$ represents the population sizes over time. The same equation holds for a subdivided population, only now $h_j(\underline{N})$ refers to the probability of non-IBD within the sub-population from which the chromosome is being sampled. For example, for the D-leut:

$$E[J_t^D \mid \underline{N}^D] = \sum_{j=0}^{t-1} h_j(\underline{N}^D) \cdot L \ , \tag{5.2}$$

where $\underline{N}^D$ denotes the leut specific population sizes over time. Consider the first observation

- that the expected numbers of junctions in chromosomes from each of the leut are very similar, for a particular model. When summing non-IBD probabilities over a fixed number of generations, small differences in population size from one leut to the next have little effect, since population size enters the calculation of the probability of non-IBD as a multiplicative term of the form $(1 - \frac{1}{2N})$. Specifically, consider a population with known probability of non-IBD at time t denoted by $h_t(\underline{N})$. Then

$$h_{t+1}(\underline{N}) = \left(1 - \frac{1}{2N_t}\right) \cdot h_t(\underline{N}) \ . \tag{5.3}$$

Now consider two possibilities for the population: either the population size at time t is N_t, or it is αN_t. The ratio of the multipliers of $h_t(\underline{N})$ in Equation 5.3 corresponding to these two possibilities is then given by

$$\frac{\left(1 - \frac{1}{\alpha 2N_t}\right)}{\left(1 - \frac{1}{2N_t}\right)} = \frac{\alpha 2N_t - 1}{\alpha 2N_t} \cdot \frac{2N_t}{2N_t - 1} = \frac{\alpha 2N_t - 1}{2N_t - 1} \cdot \frac{1}{\alpha} \sim 1 \ , \tag{5.4}$$

for large N_t and α close to 1. That is for N_t large enough, even moderate changes in population size do not affect the probability of non-IBD. The results in Table 5.6 show that for the S-leut and D-leut in particular, population sizes are large enough that the differences in population sizes between the leut result in only small differences in the expected number of junctions. For the L-leut, the very small population size at the time of subdivision results in slightly lower probabilities of non-IBD, and therefore a slightly smaller expected number of junctions.

The second observation based on Table 5.6 was that the expected number of junctions per Morgan based on model random-25 was nearly two less than that based on model random-20. This simply reflects that using a shorter generation time results in more generations over a fixed period of time. In this example, random-20 has 11 generations after founding, and random-25 has 9 generations after founding. Since the population sizes are quite large, the non-IBD probabilities are close to one, and therefore the two extra generations result in nearly two extra junctions per Morgan. In a large and young population like the Hutterites, the number of generations assumed since founding has an important effect on the expected number of junctions per Morgan. While there is a substantial difference in expectation between the two models, it is important to recognize that the variance of the number of

junctions per Morgan is large, as the simulations of Section 5.3.4 will show. In fact, there is considerable of overlap between the distributions of the number of junctions according to each model.

Simulation results

Figure 5.3 shows the mean number of junctions in both paternal and maternal chromosomes of length one Morgan, in the 27 selected Hutterites of Section 5.3.2. The horizontal axis gives the mean number of junctions observed over 100,000 simulations. There are 54 horizontal lines, one each for the maternal and paternal chromosome of the 27 individuals. The observed mean number of junctions in the chromosome is marked with a '+' along the horizontal line. The chromosomes are grouped according to leut along the vertical axis. The solid vertical lines shows the expected number of junctions, based on models random-25 (left lines) and random-20 (right lines) as shown in Table 5.6.

Figure 5.3 shows that the mean number of junctions in a chromosome, as observed over 100,000 simulations on the Hutterite pedigree, is smaller than one would expect based on either of the two random-mating populations with subdivision that we used to approximate the Hutterite's population history. To evaluate the importance of this difference, we estimated the full distribution of the number of junctions in one Morgan by simulating 5,000 random mating populations according to model random-20, and another 5,000 according to model random-25. Figure 5.4 shows the resulting simulated distributions for each of the three leut. The mean number of junctions observed over 100,000 simulations in selected Hutterite chromosomes (those shown in Figure 5.3) are marked as short vertical lines above the x-axis. The distributions under model random-20 and model random-25 are very similar, with the distribution for random-25 shifted slightly to the left. While the observed means in selected Hutterite chromosomes are smaller than the expected value of either distribution, they are well within the range of values for both models.

The smaller number of junctions observed in simulations on the Hutterite pedigree could be due to the colony structure, which is not represented in our models, or due to preferential selection of mates who are more closely related than "random" members of the population.

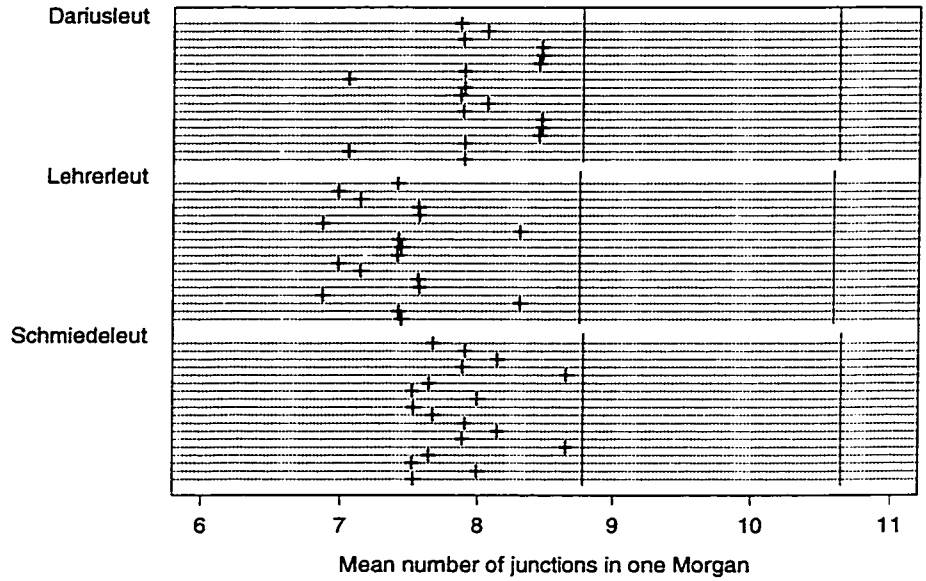
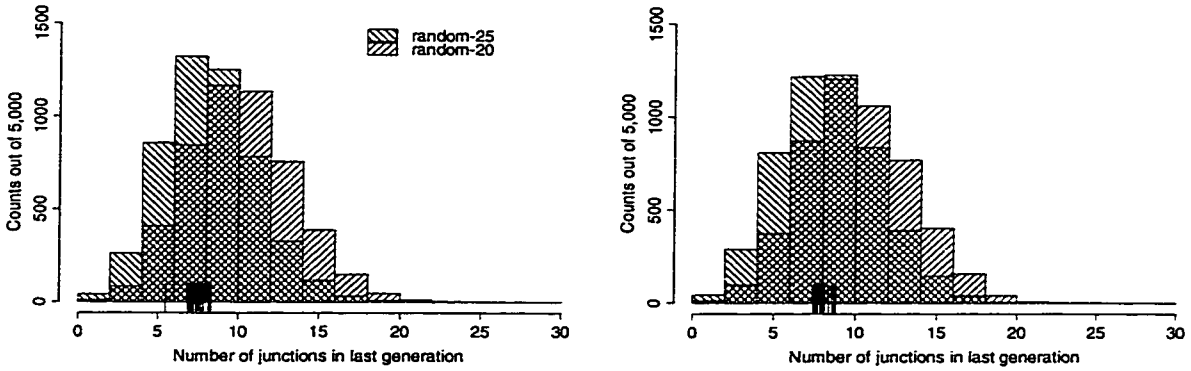
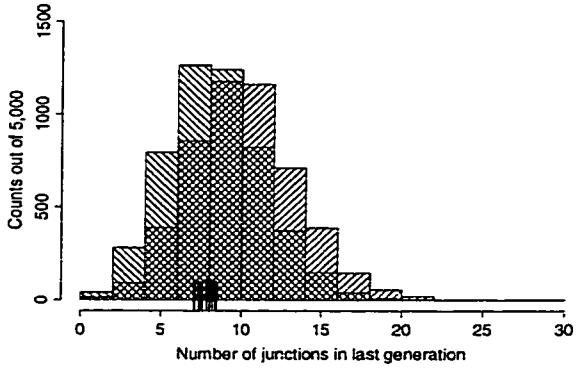


Figure 5.3: Mean number of junctions in a chromosome of length one Morgan, in both maternal and paternal chromosomes of selected Hutterites. Vertical lines represent theoretical means based on models random-25 (left lines) and random-20 (right lines).



(a) Lehrerleut

(b) Schmiedeleut



(c) Dariusleut

Figure 5.4: Distributions of the number of junctions in one Morgan, estimated by 5,000 simulations each for models random-20 and random-25. Short vertical lines above the x-axis mark the observed mean number of junctions in 18 chromosomes from the appropriate leut of the Hutterite population, based on 100,000 simulations on the Hutterite pedigree.

Table 5.7: Expected length in Morgans of an IBD tract between two randomly chosen chromosomes from the most recent generation of Hutterites, based on models random-20 and random-25.

		random-20			random-25		
		2nd csome from:					
		L-leut	S-leut	D-leut	L-leut	S-leut	D-leut
	L-leut	0.071	0.057	0.057	0.085	0.067	0.067
1st csome from:	S-leut		0.065	0.057		0.078	0.067
	D-leut			0.068			0.081

However, the large variance of the number of junctions in one Morgan implies that these results are not inconsistent with either of the random mating models we used. These results do not provide evidence of non-random mating in the Hutterite population.

5.3.5 Lengths of IBD tracts in pairs of Hutterite chromosomes

Theoretical predictions based on Hutterite population sizes

Table 5.7 shows the expected length in Morgans of an IBD tract between two randomly selected chromosomes from the most recent generation of Hutterites, based on models random-20 and random-25. The expected length is based on the approximation developed in Chapter 4. Within each of the two models, the lengths of tracts shared between chromosomes of different leut are expected to be substantially smaller than the lengths of tracts shared between chromosomes within leut. The expected lengths of a shared tract between chromosomes from different pairs of leut are the same to three decimal places. Between the two models, Table 5.7 shows that model random-25 predicts longer tracts of IBD than does model random-20.

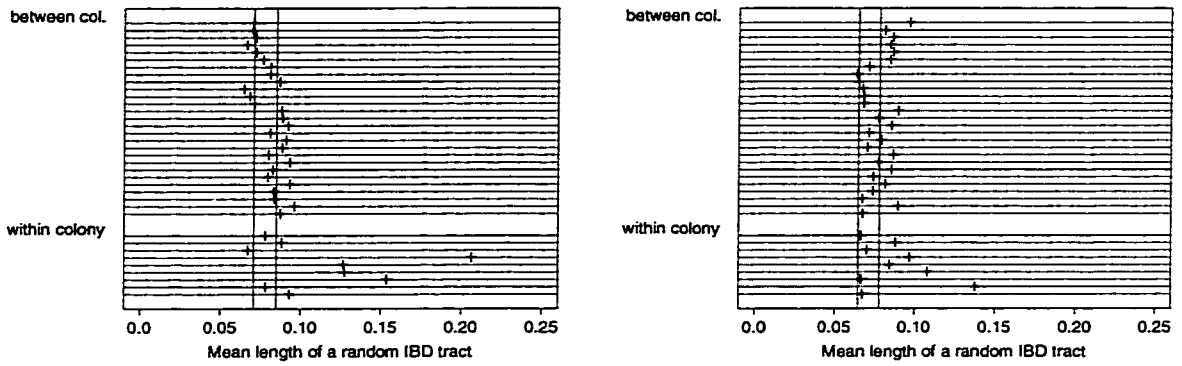
To understand the difference between the models, recall from Chapter 4 that the estimate of mean length of an IBD tract is just the expected fraction of the chromosome that is IBD, divided by the expected number of IBD tracts making up the chromosome, which is the expected number of type V junctions (Equation 4.25). In model random-25, the expected

IBD fraction is about 77% of the expected IBD fraction under model random-20, both within and between leut. The expected number of V junctions under model random-25 is about 66% of the expected number of V junctions under model random-20, both within and between leut. The increase in mean lengths under model random-25 is due to the fact that the denominator of the estimate is decreased more than the numerator of the estimate. Because the population sizes are large for both models, the expected IBD proportion is not dramatically affected by the different models. However, under model random-25, there are two fewer generations in which junctions can be formed, and so the expected number of V junctions is reduced relative to model random-20. This results in longer IBD tracts under model random-25.

Simulation results

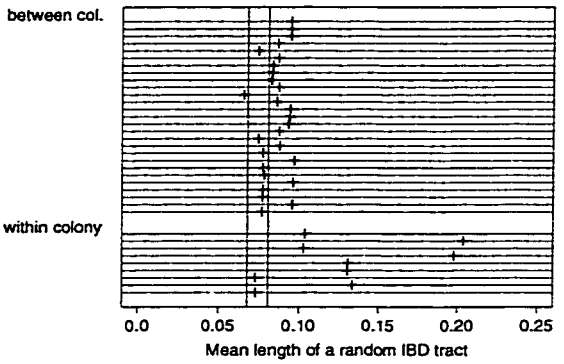
Figure 5.5 shows the mean length of a random IBD tract between each pair of the 9 paternal chromosomes of selected Hutterites within the same leut. The selection of these individuals was described in Section 5.3.2. The horizontal axis gives the mean length of a random IBD tract observed over 100,000 realizations. There are 36 horizontal lines in each plot, one each for each pair of paternal chromosomes within leut. The observed mean length of an IBD tract is marked with a '+' along the horizontal line. The chromosome pairs are grouped along the y-axis according to whether or not the individuals are in the same colony. The solid vertical lines shows the expected length of an IBD tract, based on models random-20 (left lines) and random-25 (right lines) as shown in Table 5.7. In all three leut, the observed values are quite close to those expected under models random-20 and random-25. Model random-25 in particular seems to match the data quite well. In all three leut, the longest IBD tracts were observed between members of the same colony. This is not surprising, since members of the same colony are likely to be more closely related than members of different colonies.

Figure 5.6 shows the mean length of a random IBD tract between the 9 paternal chromosomes of selected Hutterites from different leut. The horizontal axis gives the mean length of a random IBD tract observed over 100,000 realizations. There are 81 horizontal lines in



(a) Lehrerleut

(b) Schmiedeleut



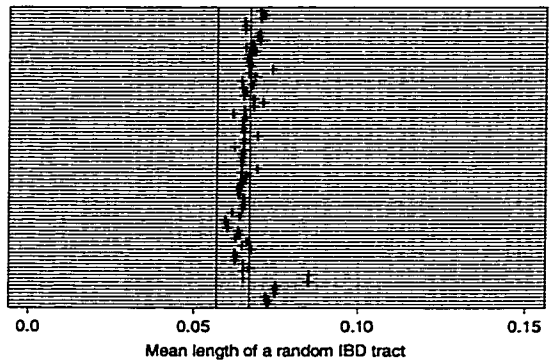
(c) Dariusleut

Figure 5.5: Mean length of a random IBD tract between the paternal chromosomes of selected Hutterites, within the same leut. Vertical lines represent theoretical means based on models random-20 (left lines) and random-25 (right lines).

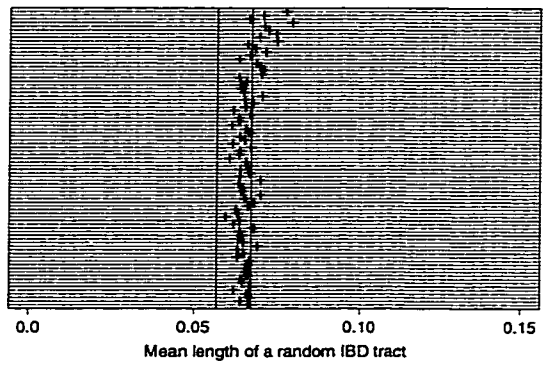
each plot, one for each pair of paternal chromosomes between two leut. The observed mean length of an IBD tract is marked with a '+' along the horizontal line. The solid vertical lines shows the expected length of an IBD tract, based on models random-20 (left lines) and random-25 (right lines) as shown in Table 5.7. Under both models, the between leut IBD tracts are shorter than the within leut IBD tracts (0.57 and 0.67 in Figure 5.6, and 0.65-0.85 in Figure 5.5). Again, this reflects the closer relationships between individuals within leut, compared to between leut. The expected values based on model random-25 seem to be in reasonable agreement with the observed data. There appears to be more variability among the mean lengths of IBD tracts between Schmiedeleut and Dariusleut chromosomes. This could be reflecting the fact that there were more interleut marriages between the Schmiedeleut and the Dariusleut (13, as opposed to 8 between the Lehrerleut and the Schmiedeleut, and 5 between the Lehrerleut and the Dariusleut). Interleut marriages lead to closer relationships between some members of the two leut than if the two leut had been completely segregated since the initial subdivision. Closer relationships result in longer lengths of IBD tracts, as seen for some Schmiedeleut-Dariusleut chromosome pairs in Figure 5.6.

Figure 5.7 shows the simulation distributions of the length of a random IBD tract between two chromosomes within the same leut, based on 5,000 realizations each under models random-20 and random-25. Vertical lines above the x-axis mark the mean length of an IBD tract estimated from 100,000 realizations, in the 36 paternal chromosome pairs from the 9 individuals selected within each leut of the Hutterite pedigree. In all three leut, the two distributions are very similar. Model random-25 has a slightly higher mean, as shown by the theoretical estimates in Table 5.7. The mean lengths observed in simulations on the Hutterite pedigree are consistent with both models.

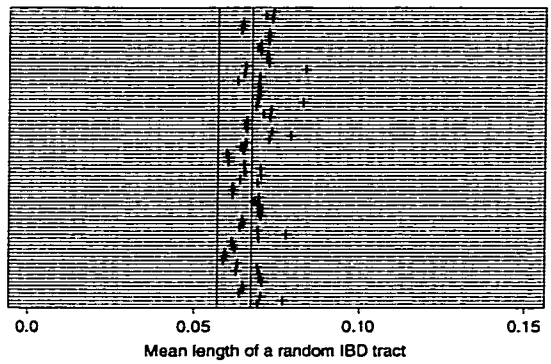
Figure 5.8 shows the corresponding simulation distributions for IBD tract lengths in chromosomes selected from different leut. The distributions are again very similar, with random-25 having a slightly higher mean. The vertical lines above the x-axis represent the observed mean lengths of IBD tracts in chromosome pairs from selected Hutterites in different leut, based on simulations on the Hutterite pedigree. As was the case for within-leut chromosome pairs, the means based on the Hutterite pedigree are consistent with either



(a) Lehrerleut-Dariusleut

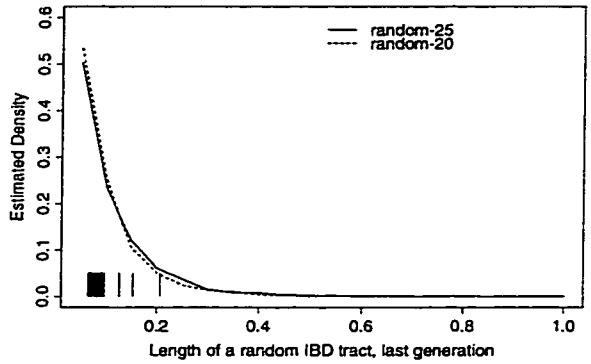


(b) Lehrerleut-Schmiedeleut

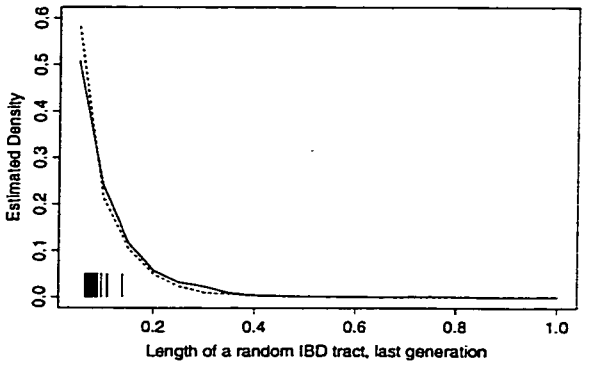


(c) Schmiedeleut-Dariusleut

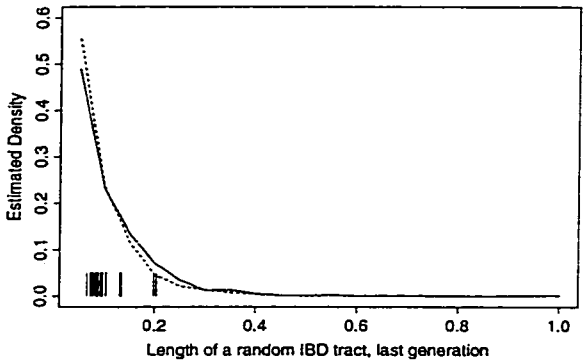
Figure 5.6: Mean length of a random IBD tract between the paternal chromosomes of selected Hutterites, between different leut. Vertical lines represent theoretical means based on models random-20 (left lines) and random-25 (right lines).



(a) Lehrerleut



(b) Schmiedeleut



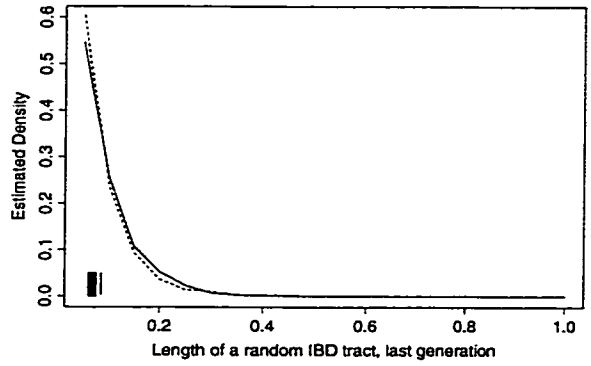
(c) Dariusleut

Figure 5.7: Distribution of the length of a random IBD tract between chromosomes within leut, estimated based on 5,000 simulations each of models random-20 and random-25. Short vertical lines above the x-axis mark the observed mean length of an IBD tract in 36 chromosome pairs from the appropriate leut of the Hutterite population, based on 100,000 simulations on the Hutterite pedigree.

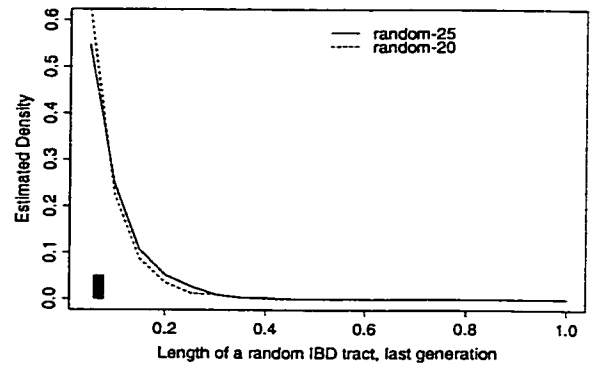
of the random mating models.

5.4 Discussion

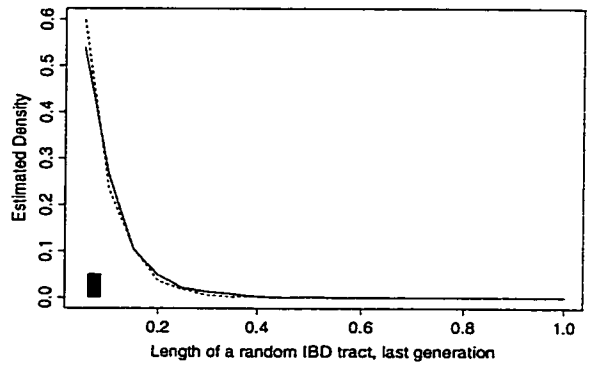
In this chapter, we first described some of the computational issues associated with simulating chromosome transmissions in random mating populations, and on pedigrees. We then reported the results of simulations on the Hutterite pedigree. For selected Hutterites, we estimated by simulation the expected number of junctions in a chromosome, and the mean length of IBD tracts shared between chromosomes from within the same colony, from different colonies with the same leut, and from different leut. The simulation results were compared to theoretical predictions of these expected values based on the application of the results of chapters 2, 3 and 4 to random mating populations intended to approximate the history and structure of the Hutterite population. We found that the average number of junctions observed in Hutterite chromosomes is smaller than that expected under either of the random mating models. This could reflect a decrease in junction formation due to the colony structure of the Hutterites, which may result in marriage between close relatives. Alternately, this could be random fluctuation about the mean, since the variance of the expected number of junctions is very high. Simulated lengths of IBD tracts were close to theoretically predicted values, although longer IBD tracts then expected were seen between chromosomes taken from the same colony.



(a) Lehrerleut-Dariusleut



(b) Lehrerleut-Schmiedeleut



(c) Schmiedeleut-Dariusleut

Figure 5.8: Distribution of the length of a random IBD tract between chromosomes chosen from different leut, estimated based on 5,000 simulations each of models random-20 and random-25. Short vertical lines above the x-axis mark the observed mean length of an IBD tract in 81 chromosome pairs from the appropriate leut-pair of the Hutterite population, based on 100,000 simulations on the Hutterite pedigree.

Chapter 6

APPLICATION OF JUNCTION THEORY TO MAP ESTIMATION

In this chapter, we apply the ideas developed in Chapter 2 to the practical problem of estimating the genetic length of a chromosomal interval using genetic data on members of the Hutterite pedigree - an isolated population whose pedigree is known. In the first section, we extend the mean, variance and covariance calculations of Chapter 2 to apply to the situation where the pedigree is known. In the second section we describe a quasi-likelihood approach to estimation, assuming that the number of junctions in the interval of interest can be observed in every sampled chromosome. Finally, we discuss some of the issues involved in the detection and resolution of junctions using IBD data on all pairs of sampled chromosomes.

6.1 Moments of the number of junctions per Morgan, conditional on pedigree structure

In this section, we develop an expression for the expected number of junctions per Morgan, conditional on the pedigree structure. We also develop approximations to the variance and covariance, and investigate their performance by simulation.

6.1.1 Mean

Let J_c be the number of junctions present on the chromosome of interest, and let j index the set of meioses which could have contributed to that chromosome. This set is made up of both meioses to all ancestors of the chromosome of interest. If the number of meioses in this set is m, let $\underline{n} = \{n_1, \ldots n_m\}$, where n_j denotes the number of junctions formed in the jth meiosis in the set. Finally, let $\mathrm{I}_c(k, j) = 1$ if the kth junction formed in the jth meiosis

is present on the chromosome of interest, and let $\mathrm{I}_c(k,j) = 0$ otherwise. Then

$$J_c \mid \underline{n} = \sum_{j=1}^{m} \sum_{k=1}^{n_j} \mathrm{I}_c(k,j) \quad . \tag{6.1}$$

Taking the expectation over the indicator function,

$$\mathrm{E}_I [J_c \mid \underline{n}] = \sum_{j=1}^{m} \sum_{k=1}^{n_j} \mathrm{E}[\mathrm{I}_c(k,j)] \quad . \tag{6.2}$$

Now $\mathrm{E}[\mathrm{I}_c(k,j)] = p_c(j)$, where $p_c(j)$ is the probability that a particular junction formed in meiosis j exists on chromosome c. This quantity is easily calculated in a recursive manner using the relation

$$p_c(j) = \frac{1}{2} p_{m(c)}(j) + \frac{1}{2} p_{p(c)}(j), \tag{6.3}$$

where $m(c)$ and $p(c)$ denote the maternal and paternal chromosomes of the individual from whom chromosome c segregated. Furthermore,

$$p_x(j) = \begin{cases} 0 & \text{if } x \text{ is a founder chromosome} \\ 1 & \text{if } x \text{ is the result of meiosis } j \\ 0 & \text{if } x \text{ is the homologue of the result of meiosis } j. \end{cases} \tag{6.4}$$

Then one can write

$$\mathrm{E}_I [J_c \mid \underline{n}] = \sum_{j=1}^{m} n_j \cdot p_c(j) \quad , \tag{6.5}$$

and if we now take the expectation over the junction formation process,

$$\mathrm{E}[J_c] = \mathrm{E}_{\underline{n}} [\, \mathrm{E}_I[J_c \mid \underline{n}] \,] = \sum_{j=1}^{m} \mathrm{E}[n_j] \cdot p_c(j) \quad . \tag{6.6}$$

We have outlined the calculation of $p_c(j)$ above, and so it remains to calculate $\mathrm{E}[n_j]$ for $j = 1, \cdots, m$. Let $H(j)$ equal the proportion of the chromosome which is not IBD in the parent of meiosis j. If recombinations happen as a Poisson process along the chromosome, this process must have rate one per Morgan, and the number of recombinations in a chromosome of length L has a Poisson distribution with mean L. A junction is just a recombination event which occurs in a region of non-IBD, and so conditional on $H(j)$, the number of junctions formed on the chromosome has a Poisson distribution with mean $H(j)L$. Thus,

$$\mathrm{E}[n_j \mid H(j)] = H(j)L,$$

and so

$$\mathrm{E}[n_j] = \mathrm{E}_H\left[\mathrm{E}[n_j \mid H(j)]\right] = \mathrm{E}[H(j)]\ L = (1 - f(j)) \cdot L \quad .$$

Here, $f(j)$ denotes the inbreeding coefficient of the parent of meiosis j, based on the pedigree structure. We can therefore calculate the expected number of junctions on chromosome c according to the equation

$$\mathrm{E}[J_c] = \sum_{j=1}^{m} (1 - f(j)) p_c(j) \cdot L \quad , \tag{6.7}$$

where L is the length of the chromosome.

6.1.2 Variance

The simulations of Section 2.1.2 show that, for a large and young randomly mating population, assuming a Poisson distribution may provide an adequate approximation for the variance of the number of junctions in a chromosome. This is convenient since the variance of a Poisson distribution is equal to its mean, and Section 6.1.1 shows that the mean is easily calculated. This motivates our consideration of the Poisson distribution for the number of junctions existing on chromosome c .

We consider the following simplifying assumptions. Suppose that

1. The number of junctions n_j formed in meiosis j has a Poisson distribution with mean $(1 - f(j))L$. It is true that *conditional* on $F(j)$, the fraction of the chromosome which is IBD with its homologue, n_j has a Poisson distribution with mean $(1 - F(j))L$. However, $F(j)$ is a random variable, and equal to $f(j)$ only in expectation. This extra variability leads to extra-Poisson variation in the distribution of n_j.

2. The number of junctions n_i formed in meiosis i is independent of the number of junctions n_j formed in meiosis j, for all $i \neq j$. This is true for meioses from unrelated individuals, and may be close to true for meioses from distantly related individuals. For meioses from a single parent to a pair of sibs (for example) it is clearly not true.

3. The presence of any one junction in chromosome c is independent of the presence of any other junction in that chromosome. That is *Pr(junction k formed in meiosis i*

exists on chromosome c | junction l formed in meiosis j exists on chromosome c)= Pr(junction k formed in meiosis i exists on chromosome c), for k, l, i and j, where $k \neq l$ *if* $i = j$.

Note that the assumptions are directly analagous to those used in Section 2.1.2 to develop the Poisson approximation to the distribution of the number of junctions in a chromosome chosen from a randomly mating population.

Let $J_c(j)$ denote the number of junctions formed in meiosis j which exist in chromosome c. Then assumption 1 implies that $J_c(j) \sim Poisson(p_c(j)(1 - f(j))L), \quad 1 \leq j \leq m$. Furthermore, assumptions 2 and 3 imply that

$$J_c(i) \text{ is independent of } J_c(j) \text{ , for } i \neq j \text{ .} \tag{6.8}$$

Therefore, under these simplifying assumptions,

$$J_c = \sum_{j=1}^{m} J_c(j) \sim Poisson\left(\sum_{j=1}^{m} p_c(j)(1 - f(j))L\right) \text{ .} \tag{6.9}$$

In the modelling application that follows, we do not use a full Poisson likelihood, but rather a quasi-likelihood in which the variance is assumed equal to the mean. The foregoing development is intended to justify that assumption.

Of the three assumptions used to develop the approximation, the third seems the least likely to be true. Particularly for the small lengths in which we are interested, the presence of one junction in the chromosome of interest tells us a lot about the presence or absence of others. For example junctions formed in the same meiosis are likely to be transmitted together, and those that are formed in different meioses are unlikely to end up on the same chromosome. To assess the effect of this assumption and the others on the variance approximation, we simulated chromosome data in 100 randomly sampled Hutterites, for a total of 200 chromosomes. Ten thousand simulations were performed, and the number of junctions in each of the 200 chromosomes was recorded. Chromosomes of length 5, 2, 1 and 0.5 cM were considered. Figure 6.1 shows the simulated variance of the number of junctions in each chromosome, plotted against the expected number of junctions in each chromosome. The dashed line indicates the line where the simulated variance is equal to the theoretical

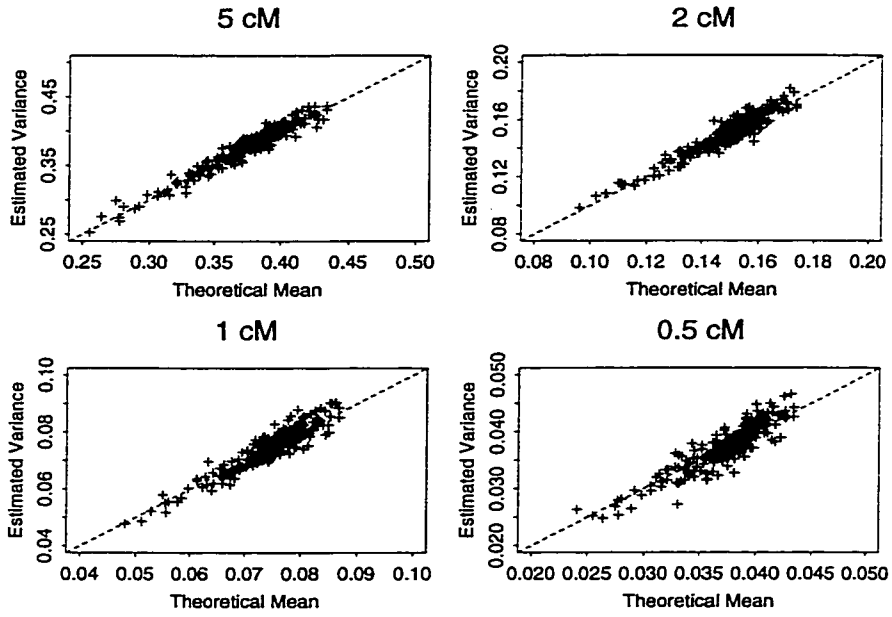


Figure 6.1: Estimated variance based on 10,000 simulations vs. theoretical mean, for 200 chromosomes from 100 randomly sampled Hutterites.

mean. The agreement between the theoretical mean and the simulated variance is quite good, for all four chromosome lengths considered. There is no evidence for a systematic departure of the variance from the mean, and as we shall see in the modelling section, this approximation works well in our estimation problem.

6.1.3 Covariance

We now consider the covariance of the numbers of junctions in two distinct chromosomes from the pedigree. Label the chromosomes c and d, and let C and D denote the sets of meioses which could have contributed to c and d respectively. If the two chromosomes are related to one another, some meioses will exist in both sets C and D. We will denote this set by CD. Meioses that contribute to c but not d are in the set $C\overline{D}$, and those that contribute to d but not c are in the set $\overline{C}D$. We can write

$$J_c \mid n_i, i \in C = \sum_{i \in C} \sum_{k=1}^{n_i} \mathrm{I}_c(k,i) = \sum_{i \in CD} \sum_{k=1}^{n_i} \mathrm{I}_c(k,i) + \sum_{j \in C\overline{D}} \sum_{l=1}^{n_j} \mathrm{I}_c(l,j) \qquad (6.10)$$

and similarly

$$J_d \mid n'_i, i' \in D = \sum_{i' \in D} \sum_{k'=1}^{n_{i'}} \mathrm{I}_d(k', i') = \sum_{i' \in CD} \sum_{k'=1}^{n_{i'}} \mathrm{I}_d(k', i') + \sum_{j' \in \overline{C}D} \sum_{l'=1}^{n_{j'}} \mathrm{I}_d(l', j') \ , \qquad (6.11)$$

where n_i denotes the number of junctions formed in meiosis i. Then

$$\begin{aligned} J_c J_d \mid n_i, i \in C \cup D \ = \ & \sum_{i \in CD} \sum_{i' \in CD} \sum_{k=1}^{n_i} \sum_{k'=1}^{n_{i'}} \mathrm{I}_c(k, i) \cdot \mathrm{I}_d(k', i') + \\ & \sum_{i \in CD} \sum_{j' \in \overline{C}D} \sum_{k=1}^{n_i} \sum_{l'=1}^{n_{j'}} \mathrm{I}_c(k, i) \cdot \mathrm{I}_d(l', j') + \\ & \sum_{j \in C\overline{D}} \sum_{i' \in CD} \sum_{l=1}^{n_j} \sum_{k'=1}^{n_{i'}} \mathrm{I}_c(l, j) \cdot \mathrm{I}_d(k', i') + \\ & \sum_{j \in C\overline{D}} \sum_{j' \in \overline{C}D} \sum_{l=1}^{n_j} \sum_{l'=1}^{n_{j'}} \mathrm{I}_c(l, j) \cdot \mathrm{I}_d(l', j') \ . \end{aligned} \qquad (6.12)$$

Now we will calculate $\mathrm{E}[J_c J_d]$, by first taking the expectation of Equation 6.12 conditional on n_i, and then taking the expectation over n_i.

Consider the last three terms of Equation 6.12. Taking the expectation with respect to the indicator variables, we need the expected value of the product of two indicator variables. For example, consider the second term of Equation 6.12. The indicator variables in this term describe (i) the event that a junction formed in a meiosis i which is in set CD exists in chromosome c and (ii) the event that a junction formed in a meiosis j which is in set $\overline{C}D$ exists in chromosome d. These events are independent, since the paths from meiosis i to chromosome c do not overlap the paths from meiosis j to chromosome d. Therefore the expectation of the product of the two indicators is equal to the product of the expectations of the indicators. Calculation of these quantities was described in the previous section. Similar arguments with respect to the independence of the indicators can be made for the third and fourth terms of Equation 6.12. Thus, taking the expectation of Equation 6.12 with respect to the indicator variables, we obtain

$$\begin{aligned} \mathrm{E}_I[J_c J_d \mid n_i, i \in C \cup D] \ = \ & \sum_{i \in CD} \sum_{i' \in CD} \sum_{k=1}^{n_i} \sum_{k'=1}^{n_{i'}} \mathrm{E}_I \left[\mathrm{I}_c(k, i) \cdot \mathrm{I}_d(k', i') \right] + \\ & \sum_{i \in CD} \sum_{j' \in \overline{C}D} \sum_{k=1}^{n_i} \sum_{l'=1}^{n_{j'}} p_c(i) \cdot p_d(j') + \end{aligned}$$

$$\sum_{j\in C\overline{D}}\sum_{i'\in CD}\sum_{l=1}^{n_j}\sum_{k'=1}^{n_{i'}} p_c(j)\cdot p_d(i') +$$
$$\sum_{j\in C\overline{D}}\sum_{j'\in \overline{C}D}\sum_{l=1}^{n_j}\sum_{l'=1}^{n_{j'}} p_c(j)\cdot p_d(j')$$
$$= \sum_{i\in CD}\sum_{i'\in CD}\sum_{k=1}^{n_i}\sum_{k'=1}^{n_{i'}} \mathrm{E}_I\left[\mathrm{I}_c(k,i)\cdot \mathrm{I}_d(k',i')\right] +$$
$$\sum_{i\in CD}\sum_{j'\in \overline{C}D} n_i n_{j'} p_c(i)\cdot p_d(j') + \sum_{j\in C\overline{D}}\sum_{i'\in CD} n_j n_{i'} p_c(j)\cdot p_d(i') +$$
$$\sum_{j\in C\overline{D}}\sum_{j'\in \overline{C}D} n_j n_{j'} p_c(j)\cdot p_d(j') \; . \qquad (6.13)$$

Now consider the first term in Equation 6.13. We can write

$$\sum_{i\in CD}\sum_{i'\in CD}\sum_{k=1}^{n_i}\sum_{k'=1}^{n_{i'}} \mathrm{E}_I\left[\mathrm{I}_c(k,i)\cdot \mathrm{I}_d(k',i')\right] = \sum_{i\in CD}\sum_{k=1}^{n_i} \mathrm{E}_I\left[\mathrm{I}_c(k,i)\cdot \mathrm{I}_d(k,i)\right] +$$
$$\sum_{i\in CD}\sum_{k=1}^{n_i}\sum_{\substack{k'=1\\k\neq k'}}^{n_i} \mathrm{E}_I\left[\mathrm{I}_c(k,i)\cdot \mathrm{I}_d(k',i)\right] + \qquad (6.14)$$
$$\sum_{i\in CD}\sum_{\substack{i'\in CD\\i\neq i'}}\sum_{k=1}^{n_i}\sum_{k'=1}^{n_{i'}} \mathrm{E}_I\left[\mathrm{I}_c(k,i)\cdot \mathrm{I}_d(k',i')\right] \; .$$

The expectation in the first term of Equation 6.14 is simply the probability that a particular junction formed in meiosis i exists in both chromosome c and chromosome d. We denote this quantity $q_{c,d}(i)$, and we defer its calculation to later in this section. To obtain expressions for the second and third terms of Equation 6.14, we assume that the indicator variables in each are independent. That is we assume (i) that the existence of junction k formed in meiosis i in chromosome c is independent of the existence of junction k' formed in meiosis i in chromosome d and (ii) that the existence of junction k formed in meiosis i in chromosome c is independent of the existence of junction k' formed in meiosis i' in chromosome d. These assumptions are similar in spirit to assumption 3 used in the variance section, although they likely closer to the truth, since they involve transmissions of junctions to different chromosomes, rather than transmissions of junctions to the same chromosome. Under these assumptions, we now have

$$\sum_{i\in CD}\sum_{i'\in CD}\sum_{k=1}^{n_i}\sum_{k'=1}^{n_{i'}} \mathrm{E}_I\left[\mathrm{I}_c(k,i)\cdot \mathrm{I}_d(k',i')\right] = \sum_{i\in CD}\sum_{k=1}^{n_i} q_{c,d}(i) +$$

$$\sum_{i\in CD}\sum_{k=1}^{n_i}\sum_{\substack{k'=1\\k\neq k'}}^{n_i} p_c(i)\cdot p_d(i) +$$
$$\sum_{i\in CD}\sum_{\substack{i'\in CD\\i\neq i'}}\sum_{k=1}^{n_i}\sum_{k'=1}^{n_{i'}} p_c(i)\cdot p_d(i')$$
$$= \sum_{i\in CD} n_i q_{c,d}(i) +$$
$$\sum_{i\in CD} n_i(n_i-1)p_c(i)\cdot p_d(i) + \qquad (6.15)$$
$$\sum_{i\in CD}\sum_{\substack{i'\in CD\\i\neq i'}} n_i n_{i'} p_c(i)\cdot p_d(i') \ .$$

Substituting this equation back into Equation 6.13 and taking the expectation over the n_i, we obtain the following:

$$\begin{aligned}\mathrm{E}[J_cJ_d] \;=\; & \sum_{i\in CD}\mathrm{E}[n_i]q_{c,d}(i) + \sum_{i\in CD}\mathrm{E}[n_i(n_i-1)]p_c(i)\cdot p_d(i) + \\ & \sum_{i\in CD}\sum_{\substack{i'\in CD\\i\neq i'}}\mathrm{E}[n_in_{i'}]p_c(i)\cdot p_d(i') + \sum_{i\in CD}\sum_{j'\in\overline{CD}}\mathrm{E}[n_in_{j'}]p_c(i)\cdot p_d(j') + \\ & \sum_{j\in\overline{CD}}\sum_{i'\in CD}\mathrm{E}[n_jn_{i'}]p_c(j)\cdot p_d(i') + \sum_{j\in\overline{CD}}\sum_{j'\in\overline{CD}}\mathrm{E}[n_jn_{j'}]p_c(j)\cdot p_d(j') \ . \qquad (6.16)\end{aligned}$$

Note that the only assumptions required for this to hold are those regarding the independence of junction existence in different chromosomes.

To complete the calculation of the covariance, we require an expression for the product of $\mathrm{E}[J_c]$ and $\mathrm{E}[J_d]$. Using Equations 6.10 and 6.11, and taking expectations first with respect to the indicator variables conditional on the n_i and then with respect to the n_i, we obtain

$$\mathrm{E}[J_c] = \sum_{i\in CD}\mathrm{E}[n_i]p_c(i) + \sum_{j\in\overline{CD}}\mathrm{E}[n_j]p_c(j) \qquad (6.17)$$

and

$$\mathrm{E}[J_d] = \sum_{i'\in CD}\mathrm{E}[n_{i'}]p_d(i') + \sum_{j'\in\overline{CD}}\mathrm{E}[n_{j'}]p_d(j') \ . \qquad (6.18)$$

Therefore

$$\mathrm{E}[J_c]\mathrm{E}[J_d] \;=\; \sum_{i\in CD}\sum_{i'\in CD}\mathrm{E}[n_i]\mathrm{E}[n_{i'}]p_c(i)p_d(i') +$$

$$\sum_{i\in CD}\sum_{j'\in\overline{C}D}\mathrm{E}[n_i]\mathrm{E}[n_{j'}]p_c(i)p_d(j') + \sum_{j\in C\overline{D}}\sum_{i'\in CD}\mathrm{E}[n_j]\mathrm{E}[n_{i'}]p_c(j)p_d(i') + \sum_{j\in C\overline{D}}\sum_{j'\in\overline{C}D}\mathrm{E}[n_j]\mathrm{E}[n_{j'}]p_c(j)p_d(j') \quad . \tag{6.19}$$

Now suppose that n_i is independent of n_j for all $i \neq j$ (assumption 2 of the variance section). Then terms 2, 3 and 4 of Equation 6.19 are equal to terms 4, 5 and 6 respectively of Equation 6.16. Therefore

$$\begin{aligned} Cov(J_c, J_c) &= \mathrm{E}[J_cJ_d] - \mathrm{E}[J_c]\mathrm{E}[J_c] = \\ &\sum_{i\in CD}\mathrm{E}[n_i]q_{c,d}(i) + \sum_{i\in CD}\mathrm{E}[n_i(n_i-1)]p_c(i)\cdot p_d(i) + \\ &\sum_{i\in CD}\sum_{\substack{i'\in CD\\ i\neq i'}}\mathrm{E}[n_i]\mathrm{E}[n_{i'}]p_c(i)\cdot p_d(i') - \sum_{i\in CD}\sum_{i'\in CD}\mathrm{E}[n_i]\mathrm{E}[n_{i'}]p_c(i)p_d(i') \\ &= \sum_{i\in CD}\mathrm{E}[n_i]q_{c,d}(i) + \sum_{i\in CD}\mathrm{E}[n_i(n_i-1)]p_c(i)\cdot p_d(i) - \\ &\sum_{i\in CD}\mathrm{E}[n_i]^2p_c(i)p_d(i) \quad . \end{aligned} \tag{6.20}$$

Finally, assume that n_i has a Poisson distribution (assumption 1 of the variance section). Then $\mathrm{E}[n_i^2] = Var[n_i] + \mathrm{E}[n_i]^2 = \mathrm{E}[n_i] + \mathrm{E}[n_i^2]$, and so $\mathrm{E}[n_i(n_i-1)] = \mathrm{E}[n_i^2]$. Therefore

$$Cov(J_c, J_d) = \sum_{i\in CD}\mathrm{E}[n_i]q_{c,d}(i) \quad . \tag{6.21}$$

This expression is equal to the expected number of junctions shared between chromosomes c and d, and we will denote it by $\mathrm{E}[S_{c,d}]$.

Calculation of $\mathrm{E}[S_{c,d}]$

Recall that

$$\mathrm{E}[S_{c,d}] = \sum_{i\in CD}\mathrm{E}[n_i]q_{c,d}(i) \ ,$$

where CD is the set of meioses that can contribute to both chromosomes c and d. We showed in section 6.1.1 that $\mathrm{E}[n_i] = (1 - f(i))L$, and so we need only calculate $q_{c,d}(i)$, the probability that a particular junction formed in meiosis i exists on both chromosomes c and

d. This is done in a recursive manner similar to the calculation of $p_c(i)$, the probability that a particular junction formed in meiosis i exists on chromosome c.

First note that $q_{x,y}(i) = q_{x,y}(i)$ for all chromosomes x and y, and meioses i. Also,

$$q_{x,y}(i) = \frac{1}{2}q_{m(x),y}(i) + \frac{1}{2}q_{p(x),y}(i) \tag{6.22}$$

where $m(x)$ and $p(x)$ denote the maternal and paternal chromosomes in the parent of chromosome x, as long as x is not an ancestor of y. This equation can be applied recursively, using the following conditions:

$$q_{x,y}(i) = \begin{cases} 0 & \text{if either } x \text{ or } y \text{ is a founder chromosome} \\ 0 & \text{if } x \text{ is the product of meiosis } i, \text{ and } y \text{ is the homologue of } x \\ p_x(i) & \text{if } x = y \\ p_y(i) & \text{if } x \text{ is the product of meiosis } i \\ 1 & \text{if } x = y \text{ and } x \text{ is the product of meiosis } i \ . \end{cases} \tag{6.23}$$

So we can calculate

$$\mathrm{E}[S_{c,d}] = \sum_{i \in CD} (1 - f(i))\, q_{c,d}(i) L \ , \tag{6.24}$$

Under the conditions which imply $Var[J_c] = \mathrm{E}[J_c]$, we also have $Cov(J_c, J_d) = \mathrm{E}[S_{c,d}]$. Using the simulations described in Section 6.1.2, we can investigate how well this approximation to the covariance works. Figure 6.2 shows the simulated covariance of the numbers of junctions in each pair of chromosomes, plotted against the expected number of junctions shared between that pair. The dashed line indicates the line where the simulated covariance is equal to the expected number shared. The clusters of points in the plots show the variety of relationships between pairs of chromosomes in the sample. Most are distantly related and thus have a very small covariance. Negative estimated covariances demonstrate that even 10,000 simulations is not enough to estimate very small covariances. The clusters with higher observed covariances represent more closely related chromosomes. The approximation of the covariance by the expected number shared seems to be adequate, although it may be an underestimate for very closely related individuals.

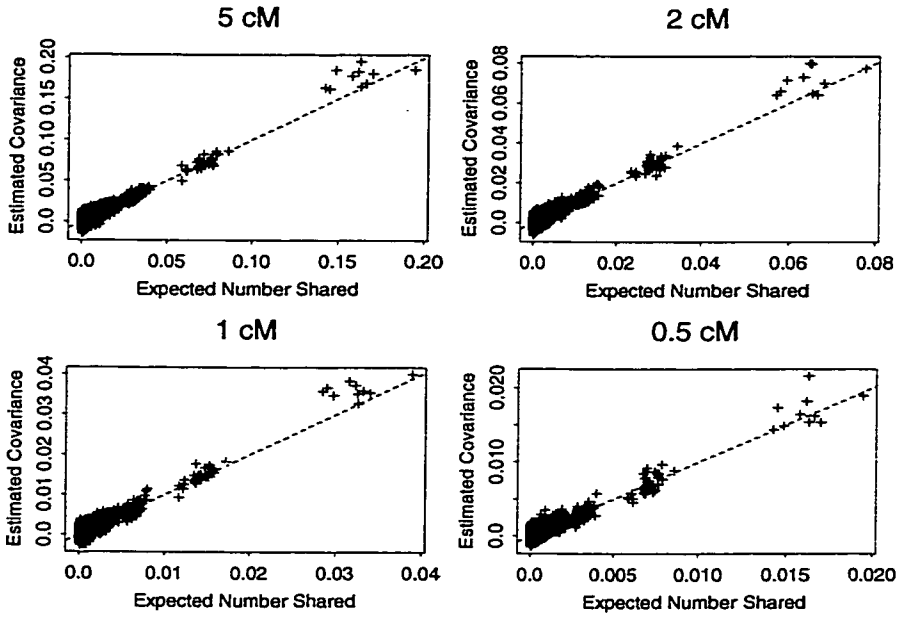


Figure 6.2: Estimated covariance based on 10,000 simulations vs. expected number of junctions shared, for all distinct pairs of 200 chromosomes from 100 randomly sampled Hutterites.

6.2 Quasi-likelihood modelling of junction counts

In this section, we describe a quasi-likelihood approach to estimating the length d of an interval (in Morgans), using the theoretical results of the previous section. We assume that the data consist of counts of the number of junctions in the interval of interest on each of a sample of chromosomes, which will come in pairs within individuals. We further assume that we know the pedigree relating all the chromosomes, going back to the founders of the population. Note that this requires we be able to distinguish the maternal and paternal chromosomes within an individual. We first discuss estimation under the assumption that the data are independent. We find that the variance of the estimate is badly underestimated in this case, and so we next use the covariance approximation developed in the Section 6.1.3.

6.2.1 *Independence*

Let $J = (J_1, J_2, \cdots, J_n)^T$ denote the number of junctions in the interval of interest in n Hutterite chromosomes. Let $\mu = \mathrm{E}[J]$ and let V be the variance -covariance matrix of J. Then $\mu_i = X_i d$, where d is the length of the interval in Morgans, and $X_i = \sum_{j=1}^{m}(1 - f(j))p_i(j)$, where j indexes the set of all meioses that could have contributed to chromosome i. Under the Poisson approximation discussed earlier, and assuming that J_i is independent of J_j for all $i \neq j$, $V = d \cdot diag\{X\}$. Letting $D = \frac{d\mu}{dd} = X$ we can form the quasi-score equation

$$\begin{aligned} U(d) &= D^T V^{-1}(J - \mu) \\ &= X^T diag\left\{\frac{1}{X}\right\}\frac{1}{d}(J - Xd) \\ &= \frac{1}{d}1^T(J - Xd) \\ &= \frac{1}{d}\left(\sum_{i=1}^{n} J_i - d\sum_{i=1}^{n} X_i\right) \end{aligned} \tag{6.25}$$

The quasi-score has the following properties (see, for example, [23]):

$$\mathrm{E}[U(d)] = 0 \tag{6.26}$$

$$Cov[U(d)] = D^T V^{-1} D = i_d \tag{6.27}$$

$$-\mathrm{E}\left[\frac{dU(d)}{dd}\right] = D^T V^{-1} D\ , \tag{6.28}$$

which motivate estimation and inference on d as if 6.25 were actually the score function from an ordinary likelihood. We therefore estimate d by setting $U(\hat{d}) = 0$ and solving for $\hat{d}$, and estimate its variance by $i_d^{-1} = \left(D^T V^{-1} D\right)^{-1}$. Setting Equation 6.25 equal to zero and solving for $\hat{d}$, we obtain

$$\hat{d} = \frac{\sum_{i=1}^{n} J_i}{\sum_{i=1}^{n} X_i} \tag{6.29}$$

The variance of the estimate is estimated by

$$\begin{aligned} Var(\hat{d}) = i_d^{-1} &= \left(D^T V^{-1} D\right)^{-1} \\ &= \left(\frac{1}{d}X^T diag\left\{\frac{1}{X}\right\}X\right)^{-1} \\ &= \frac{d}{\sum_{i=1}^{n} X_i}\ . \end{aligned} \tag{6.30}$$

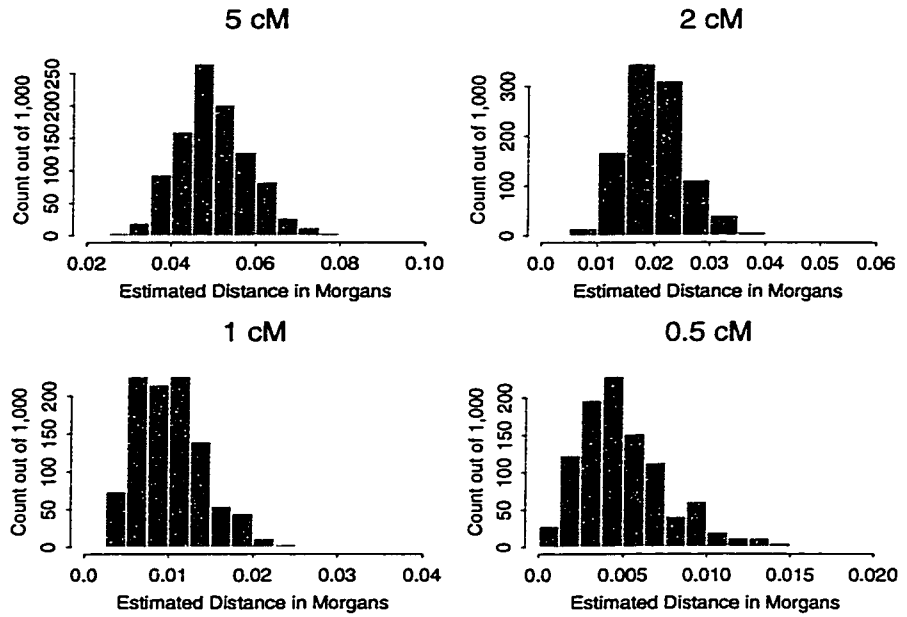


Figure 6.3: Simulation distributions of estimates of interval length based on the independence assumption, for true values of $d = 5, 2, 1$ and 0.5 cM. Data used were 200 chromosomes from 100 randomly sampled Hutterites.

Under appropriate limiting conditions (see [23]), $\hat{d}$ is approximately unbiased for d, and asymptotically Normal with limiting variance i_d^{-1}.

We tested this approach by simulating data on the Hutterite pedigree. A random sample of one hundred Hutterites who were alive at the time of the 1981 census was chosen, for a total of 200 chromosomes ($n = 200$). Chromosome data in these individuals were simulated using the entire ancestry, as discussed in Chapter 5. Once the chromosomes had been simulated, the number of junctions in the interval of interest was counted, in each of the 200 chromosomes. Estimates of $\hat{d}$ and $Var(\hat{d})$ were calculated as described above. Intervals of length 5 cM, 2 cM, 1 cM and 0.5 cM were considered, and 1000 simulations were done.

Figure 6.3 shows the simulation distributions of the estimates obtained for each of the four interval lengths. For the longer intervals the distributions are symmetric, while they are slightly skewed for the shorter intervals. Table 6.1 summarizes the results of the simulations. Rows show (i) mean of $\hat{d}$ over 1,000 simulations; (ii) standard deviation of $\hat{d}$ over 1,000 simulations; and (iii) mean over 1,000 simulations of the model based estimate of

Table 6.1: Results of 1,000 simulations estimating d assuming independence and using 200 chromosomes from 100 randomly sampled Hutterites.

	True interval length (M)			
	0.05	0.02	0.01	0.005
(i) mean $\hat{d}$	0.0499	0.0199	0.0101	0.0050
(ii) std. dev. $\hat{d}$	0.0085	0.0054	0.0040	0.0028
(iii) mean $\sqrt{Var(\hat{d})}$	0.0058	0.0036	0.0026	0.0018

the standard deviation of $\hat{d}$. As the histograms showed, $\hat{d}$ is estimating d well. However, rows (ii) and (iii) show that this quasi-likelihood model is dramatically underestimating the variance of the estimate (since the simulation based estimate in row (ii) is so much larger than the model based estimate in row(iii)). This underestimation of the variance could be due to correlation between chromosomes, which we had assumed to be independent. This motivates our use of E[S] as an approximation to the covariance.

6.2.2 Covariance approximated by E[S]

The quasi-likelihood model described in the previous section can be applied in exactly the same manner when the variance matrix is not diagonal, although the conditions required for consistency and asymptotic normality of the estimate are more complicated (see [23]). When we approximate $Cov(J_i, J_j)$ by $\mathrm{E}[S_{i,j}] = \sum_{k \in IJ} (1 - f(k))\, q_{i,j}(k) d = Y_{i,j} d$, the variance-covariance matrix V is given by

$$V = d \begin{bmatrix} X_1 & Y_{1,2} & \cdots & Y_{1,n} \\ Y_{1,2} & X_2 & \cdots & Y_{2,n} \\ & & & \\ Y_{1,n} & Y_{2,n} & \cdots & X_n \end{bmatrix}.$$

The quasi-score matrix is formed in exactly the same way, and letting $W^T = d \cdot D^T V^{-1}$ for notational convenience, we obtain

$$U(d) \;=\; D^T V^{-1} (J - \mu)$$

$$
\begin{aligned}
&= \frac{1}{d} W^T (J - Xd) \\
&= \frac{1}{d} \left(\sum_{i=1}^{n} W_i J_i - d \sum_{i=1}^{n} W_i X_i \right) . \qquad (6.31)
\end{aligned}
$$

Notice that the weights W depend only on X and $Y_{i,j}$, the components of the covariance matrix. Setting Equation 6.31 equal to zero and solving for $\hat{d}$, we obtain

$$\hat{d} = \frac{\sum_{i=1}^{n} W_i J_i}{\sum_{i=1}^{n} W_i X_i} \qquad (6.32)$$

The variance of the estimate is estimated by

$$
\begin{aligned}
Var(\hat{d}) = i_{\hat{d}}^{-1} &= \left(D^T V^{-1} D \right)^{-1} \\
&= \left(\frac{1}{d} W^T X \right)^{-1} \\
&= \frac{d}{\sum_{i=1}^{n} W_i X_i} . \qquad (6.33)
\end{aligned}
$$

We applied this model to the simulated data described in the previous section.

Figure 6.4 shows the simulation distributions of the estimates obtained for each of the four interval lengths. Again, the distributions are symmetric for longer intervals, and more skewed for shorter intervals. Table 6.2 summarizes the results of the simulations. Rows show (i) mean of $\hat{d}$ over 1,000 simulations; (ii) standard deviation of $\hat{d}$ over 1,000 simulations; and (iii) mean over 1,000 simulations of the model based estimate of the standard deviation of $\hat{d}$. Again, $\hat{d}$ is estimating d well. The close agreement between rows (ii) and (iii) shows that incorporating an approximation to the covariance of the observations into the model dramatically improved our variance estimate.

6.2.3 Effects of sample composition on the variance of the estimate

The results presented in the previous section are all the result of application of the model to a data set which is made up of 200 chromosomes from 100 randomly sampled Hutterites. The availability of the pedigree of the Hutterites makes it possible to consider different sampling schemes - for example sampling groups of relatives. In addition to the randomly sampled group, we considered two other data sets of close to 100 people from the Hutterite population. One alternate data set consisted of groups of first cousins. First we selected

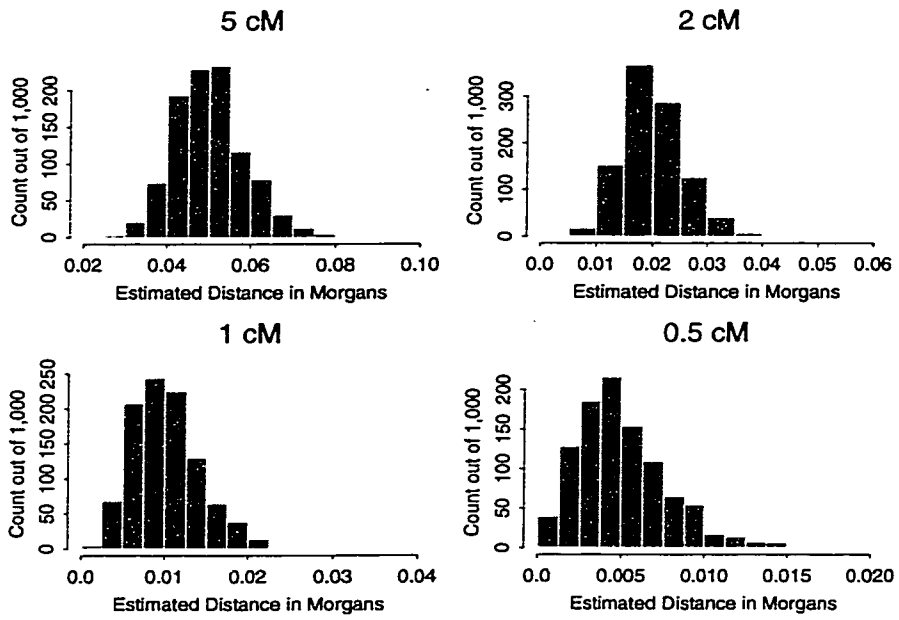


Figure 6.4: Simulation distributions of estimates of interval length using $Cov(J_c, J_d) = \mathrm{E}[S_{c,d}]$, for true values of $d = 5, 2, 1$ and 0.5 cM. Data used were 200 chromosomes from 100 randomly sampled Hutterites.

Table 6.2: Results of 1,000 simulations estimating d using $Cov(J_c, J_d) = \mathrm{E}[S_{c,d}]$ and 200 chromosomes from 100 randomly sampled Hutterites.

	True interval length (M)			
	0.05	0.02	0.01	0.005
(i) mean $\hat{d}$	0.0500	0.0199	0.0101	0.0050
(ii) std. dev. $\hat{d}$	0.0083	0.0053	0.0039	0.0027
(iii) mean $\sqrt{Var(\hat{d})}$	0.0085	0.0054	0.0038	0.0026

one individual from each leut. For each of these individuals, we chose four maternal cousins and four paternal cousins. These were chosen so that none of the cousins were siblings - this can be done in the Hutterites because couples have so many children that for each person, there are many "cousinships" to chose from. For each of these eight cousins of the proband, we selected four more cousins who were not siblings, and were not cousins of the original proband. For example, if we call the original proband A, we included four of his maternal cousins, one of whom was labeled B. We then included four paternal cousins of B, who are not cousins of A unless a sib-exchange marriage took place. This was in fact the case in one instance, and so the sample ended up consisting of 97 people.The second alternate data set was made up of 27 sibships. The 27 probands were those used in Chapter 5 - there were three individuals from three colonies in each of the three leut. The individuals were randomly chosen from the colonies, and the colonies were somewhat closely related, although they were not sister colonies. For each of these 27 probands, three siblings were included in the sample. Some individuals did not have three siblings, and so the total number of individuals in the data set was 96.

The properties of the quasi-likelihood estimate were assessed for each of the alternate data sets by simulation. Simulations consisted of 1,000 iterations, and interval lengths of 5, 2, 1 and 0.5 cM were considered. Table 6.3 shows the mean of the estimate $\hat{d}$ over all simulations for each of the data sets and each interval length. The estimate $\hat{d}$ appears to be unbiased at all lengths considered, for each of the data sets. Table 6.3 also shows the standard deviation of the estimate $\hat{d}$, and this is where differences between the data sets are apparent. The estimate has a substantially larger standard deviation in the cousinship data as compared to the random sample, and a still larger standard deviation in the data set consisting of sibships. This is explained by the fact that in the data sets consisting of close relatives, the correlation between observations is much larger, and this leads to an increase in the variance of the estimate. In effect, sampling close relatives is like sampling substantially less than 100 people. The correlations between groups of sibs will be larger than those between groups of cousins, which explains why the standard deviation of the estimate based on the sibship data set is higher again than that based on cousinships.

These results might lead one to conclude that it is best to sample completely unrelated

Table 6.3: Simulation results comparing estimation of $\hat{d}$, using different data sets.

		True interval length in Morgans (d)			
	Data set	0.05	0.02	0.01	0.005
mean $\hat{d}$	100 random	0.0500	0.0199	0.0101	0.0050
	cousinships	0.0501	0.0198	0.0101	0.0050
	27 sibships	0.0501	0.0201	0.0101	0.0052
std. dev. $\hat{d}$	100 random	0.0083	0.0053	0.0039	0.0027
	cousinships	0.0093	0.0057	0.0042	0.0030
	27 sibships	0.0104	0.0067	0.0048	0.0035

individuals, since observations on their chromosomes will be independent. This would be true if we could in fact count the number of junctions in an interval on a chromosome. However, in any application to real data, junctions are detected by a change in IBD state between two chromosomes. Unrelated chromosomes will be non-IBD along their entire length, and therefore none of the junctions contained in those chromosomes will be visible. We discuss the issues of junction detection and resolution in some detail in the next section.

6.3 Detection and resolution of junctions using continuous IBD data

The results of Section 6.2 demonstrate that there is substantial information in Hutterite chromosomes for estimating small genetic distances. In this section, we examine how well this information can be extracted, assuming that we have the most informative data possible - continuous IBD data on all pairs of chromosomes in the sample.

Figure 6.5 shows the four kinds of chromosomes that exist, considering the neighbourhood of the chromosome immediately surrounding the junction of interest. Different shadings represent different ancestral types. Chromosome 1 carries the junction; chromosome 2 carries one ancestral type; chromosome 3 carries the other ancestral type, and chromosome 4 carries neither. Consider the IBD information one would obtain from comparing pairs of these different types of junctions. Comparing a given type with another of the same

Figure 6.5: Four kinds of chromosome with respect to junction depicted. Different shadings represent different ancestral types.

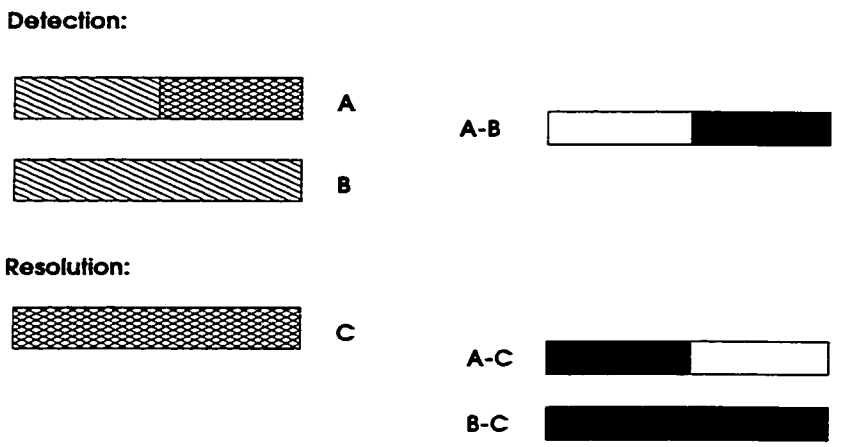


Figure 6.6: Detection and resolution of a junction using continuous IBD data. On the left, different shadings represent different ancestral types, and on the right, white represents IBD and black non-IBD.

type will result in continuous IBD - there is no information about junctions here. Similarly, comparing type 4 with any of the other three types, or comparing type 2 with type 3, will result in continuous non-IBD. Thus information about the existence of the junction comes from comparison of chromosome types 1 and 2, or types 1 and 3. Both pairs show a change in IBD status, which indicates the presence of a junction (the junction is *detected*). It is not apparent from a single such pair which chromosome carries the junction and which one does not. In order to determine which chromosome carries the junction (to *resolve* the junction), there must be a third chromosome in the sample which has the ancestor on the side of the junction which is not shared by the first two. Figure 6.6 shows the detection and resolution of a junction using continuous IBD data. The IBD information between

Table 6.4: Simulation results: average number of junctions existing, average proportion detected, average proportion resolved in 200 chromosomes from 100 randomly sampled Hutterites.

d	Number existing	Proportion detected	Proportion resolved
5 cM	52.69	0.94	0.75
2 cM	21.14	0.94	0.75
1 cM	10.74	0.94	0.74
0.5 cM	5.26	0.94	0.75

chromosomes A and B indicates that a junction exists (it is detected). However, from this data alone, it is unclear whether the junction is on chromosome A or B. When the IBD information between chromosomes A and C is considered, a similar conclusion is reached. Thus, either the junction exists on chromosome A alone, or the junction exists on both B and C. Consideration of the IBD data between chromosomes B and C indicates that they cannot both have the junction - since there is continuous non-identity in the region. Thus the junction exists on chromosome A (it is resolved). Thus, given that a chromosome of type 1 exists in the sample of chromosomes, the sample must include at least one of type 2 or type 3 in order to detect the junction, and at least one of each to resolve the junction. The quasi-likelihood model which appropriately models the covariance structure requires that all junctions be resolved, since it requires junction counts in individual chromosomes.

A simulation study was carried out to assess how well junctions are detected and resolved in the data set consisting of chromosomes from 100 randomly chosen Hutterites. The simulations are the same ones used to assess the quasi-likelihood model in Section 6.2. For each of the 1,000 iterations, we noted the number of junctions existing in the interval of interest in the 200 chromosomes, the number of junctions which were detectable using continuous IBD data, and the number of junctions which were resolvable using continuous IBD data. Table 6.4 shows the average number of junctions existing in the sample, the average proportion of those which were detectable, and the average proportion of detectable junctions which were resolvable. The results show that in this sample, 94% of existing

junctions can be detected, and about 75% of those can be resolved. Numbers this high are due to the history of the Hutterite population, which is ideal for junction detection and resolution. The Hutterites are a young population, and they have grown very rapidly. This means that for a junction existing in the sample, it is likely that both ancestral types still persist, and therefore the junction should be detectable and resolvable.

6.3.1 Characteristics of undetected or unresolved junctions

Because the detection of an existing junction requires existence in the sample of at least one copy of one of the ancestral types, it follows that older junctions may be less likely to be detected and in turn resolved. This is because for older junctions, there has been more opportunity for one or both ancestral types to become extinct - at least in the set of sampled chromosomes. To explore this possibility, some simulations were conducted in which a surrogate for the time of formation of each junction was recorded. The surrogate used was the year of birth of the individual who gave rise to the meiosis in which the junction was first formed. It would be more accurate to consider the year of birth of the individual in whose chromosome the junction first existed, but this surrogate seems adequate for our purposes. Junctions existing in the sample of 200 Hutterite chromosomes were then divided into bins according to ten year "formation cohorts", and the number of junctions, proportion detected, and proportion of observed junctions that were resolved were recorded in each bin, for each simulation. Simulations were performed for an interval length of 50 cM, in order to ensure a reasonable number of junctions would fall into each formation cohort. Simulation of an interval of this length is more computationally demanding, so only 500 iterations were used.

Table 6.5 shows the mean number of junctions existing in the sample, the mean proportion detected, and the mean proportion resolved, by formation cohort, for the data set consisting of 200 chromosomes from 100 random Hutterites. The proportion detected is close to 100% until as far back as the early 1800s, where it drops quite quickly. The proportion resolved follows a similar pattern, although the best resolution rate is just over 80%. These observations confirm that older junctions are less likely to be detected and

Table 6.5: Simulation results: average number of junctions existing, average proportion of junctions detected and average proportion of detected junctions that are resolved, by formation cohort, estimated by 500 simulations over an interval of 50 cM. (200 chromosomes from 100 randomly sampled Hutterites).

Formation cohort	Number existing	Proportion detected	Proportion resolved
1690-1700	0.66	0.15	0.00
1700-1710	2.61	0.52	0.00
1710-1720	5.25	0.83	0.45
1720-1730	9.63	0.76	0.31
1730-1740	5.74	0.78	0.32
1740-1750	9.16	0.68	0.40
1750-1760	14.55	0.77	0.51
1760-1770	10.97	0.78	0.50
1770-1780	19.50	0.84	0.59
1780-1790	9.85	0.89	0.61
1790-1800	24.99	0.91	0.70
1800-1810	14.83	0.93	0.71
1810-1820	28.94	0.95	0.70
1820-1830	23.43	0.94	0.69
1830-1840	24.05	0.99	0.81
1840-1850	24.84	0.97	0.78
1850-1860	31.33	0.99	0.81
1860-1870	31.01	0.98	0.82
1870-1880	32.57	0.99	0.82
1880-1890	27.85	0.97	0.80
1890-1900	34.61	0.99	0.84
1900-1910	26.17	0.99	0.82
1910-1920	38.25	0.98	0.81
1920-1930	24.60	1.00	0.86
1930-1940	20.79	0.99	0.82
1940-1950	26.83	0.99	0.82
1950-1960	8.74	0.99	0.84

Table 6.6: Simulation results: average number of junctions existing, average proportion detected, average proportion resolved, over 1,000 simulations of an interval of length 5 cM.

Data set	Number existing	Proportion detected	Proportion resolved
100 random	52.69	0.94	0.75
98 distant	55.37	0.94	0.76
97 in cousinships	46.88	0.93	0.73
96 in 27 sibships	33.83	0.90	0.67

subsequently resolved.

6.3.2 Effects of sample composition on detection and resolution

Given a particular junction that exists in the sample, sampling other individuals who are the direct descendants of the individual from whom the junction originated should improve the probabilities of detecting and resolving that junction. For example, sampling cousins should optimize for detection and resolution of junctions formed in meioses from their grandparents to their parents, while sampling siblings should optimize detection and resolution of junctions newly formed in that generation. This implies that different samples will give rise to different detection and resolution probabilities, and that it may be possible to optimize the sample in this respect.

Table 6.6 shows the overall detection and resolution rates as estimated by simulation, for three alternative samples of approximately 100 individuals. The random sample is included for reference, and the composition of the cousinship and sibship samples was described in Section 6.2.3. The data set labelled "distant" consists of one individual sampled from each of a set of colonies. The colonies were chosen to be as distantly related as possible - the approximate approach was to chose colonies which initially formed prior to 1940. The intent of this approach was to sample very distantly related Hutterites, in hopes of detecting some of the older junctions which are not as well detected by the random sample of Hutterites. The simulations show that detection and resolution were no better in the distant data set. Detection and resolution were worse in the data sets containing more closely related

individuals, and in particular in the data set consisting of 27 sibships.

To explore how these differences were related to junction age, we recorded detection and resolution proportions in each formation cohort, over 500 simulations of an interval of length 50 cM in the three additional data sets. Table 6.7 shows the mean proportion of existing junctions that were detected, and the mean proportion of detected junctions that were resolved, by formation cohort, for the four data sets. Detection and resolution proportions are nearly identical between the distant data set and the random data set, in all formation cohorts. The detection proportions differ by a maximum of 0.02, and while there is more variability in the resolution proportions, there is no consistent pattern of difference. This result suggests that choosing individuals from long established colonies does not necessarily result in distantly related individuals. It is also possible that the random data set achieves the maximum detection and resolution proportions possible in a sample of about 100 Hutterites.

Detection proportions in the data set consisting of cousinships are as good as those in the random data set for recently formed junctions, in that they are nearly 100%. However, the detection proportions are consistently worse for older junctions. The resolution proportions for the cousinship data set seem to be slightly better for recently formed junctions, but again older junctions are less well resolved by this data set. These observations are consistent with the idea that sampling cousins should improve detection and resolution of junctions formed in meioses to their parents.

Finally, junction detection in the sibship data set is as good as that in the random dataset only for very recently formed junctions. The detected proportion drops off more quickly (as junctions get older) than in the random data set. Junctions formed very recently are much more likely to be resolved using the sibship data set, but moderately old junctions (e.g. 1780-1890) are much less likely to be resolved in the sibships. These observations are consistent with the idea that sampling siblings should improve detection and resolution of junctions formed in the most recent generation. Unfortunately, this improvement comes with a cost of reduced detection and resolution of older junctions.

Table 6.7: Simulation results: average proportion of junctions detected and average proportion of detected junctions that are resolved, by formation cohort and data set, estimated by 500 simulations over an interval of 50 cM.

Formation	Proportion detected				Proportion resolved			
cohort	random	distant	cousins	siblings	random	distant	cousins	siblings
1690-1700	0.15	0.13	0.12	0.06	0.00	0.00	0.00	0.00
1700-1710	0.52	0.53	0.49	0.44	0.00	0.00	0.00	0.00
1710-1720	0.83	0.82	0.81	0.74	0.45	0.48	0.40	0.32
1720-1730	0.76	0.75	0.70	0.65	0.31	0.31	0.25	0.23
1730-1740	0.78	0.78	0.70	0.63	0.32	0.34	0.27	0.22
1740-1750	0.68	0.67	0.60	0.58	0.40	0.38	0.36	0.29
1750-1760	0.77	0.77	0.74	0.68	0.51	0.53	0.46	0.39
1760-1770	0.78	0.79	0.73	0.68	0.50	0.51	0.45	0.40
1770-1780	0.84	0.84	0.81	0.75	0.59	0.59	0.51	0.41
1780-1790	0.89	0.89	0.84	0.79	0.61	0.61	0.51	0.45
1790-1800	0.91	0.92	0.87	0.82	0.70	0.70	0.61	0.50
1800-1810	0.93	0.94	0.91	0.86	0.71	0.72	0.64	0.52
1810-1820	0.95	0.95	0.91	0.88	0.70	0.70	0.62	0.52
1820-1830	0.94	0.93	0.91	0.87	0.69	0.70	0.61	0.49
1830-1840	0.99	0.99	0.97	0.92	0.81	0.81	0.69	0.57
1840-1850	0.97	0.97	0.96	0.92	0.78	0.77	0.68	0.59
1850-1860	0.99	0.99	0.97	0.93	0.81	0.82	0.71	0.58
1860-1870	0.98	0.98	0.97	0.94	0.82	0.83	0.76	0.64
1870-1880	0.99	0.99	0.98	0.92	0.82	0.82	0.76	0.57
1880-1890	0.97	0.97	0.95	0.93	0.80	0.81	0.77	0.65
1890-1900	0.99	0.99	0.98	0.95	0.84	0.84	0.79	0.65
1900-1910	0.99	0.99	0.99	0.96	0.82	0.83	0.86	0.66
1910-1920	0.98	0.96	0.99	0.93	0.81	0.83	0.85	0.70
1920-1930	1.00	0.98	0.99	0.99	0.86	0.82	0.85	0.86
1930-1940	0.99	0.99	0.99	1.00	0.82	0.84	0.86	0.94
1940-1950	0.99	0.99	0.99	1.00	0.82	0.83	0.84	0.93
1950-1960	0.99	0.99	0.99	0.99	0.84	0.82	0.86	0.85

6.4 Discussion

In this chapter, we developed a method for estimating small map distances, using counts of numbers of junctions existing in Hutterite chromosomes. In order to do this, we first extended the mean, variance and covariance calculations of Chapter 2 to be applicable to a population for which the pedigree is known. We then described a quasi-likelihood model for the junction counts, assuming independence of counts in different chromosomes. The estimate of distance based on this model appeared to be unbiased, but the model based variance of the estimate was too small. We corrected this problem using an approximation to the covariance of the number of junctions in different chromosomes. The corrected model gave good variance estimates, and for a sample of only 200 Hutterite chromosomes, the standard error on an estimate of a map distance of 0.005 M was 0.0027 M. This demonstrates that there is substantial information contained in Hutterite chromosomes for the estimation of small map distances. We also considered issues of junction detection and resolution using continuous IBD data on all pairs of chromosomes in the sample. In particular, we found that older junctions are more difficult to detect and resolve, and that sample composition can affect overall detection and resolution rates.

Chapter 7

DISCUSSION

In this dissertation, we have developed an approach which allows us to study the effects of some aspects of population history and structure on the number of ancestral segments expected in a chromosome sampled from a modern population. We also modelled the lengths of IBD segments, and considered an application of the theory to the estimation of small map distances using the Hutterite pedigree.

In Chapter 2, we show that for very small populations, different growth patterns can result in dramatically different numbers of ancestral segments. However, different growth patterns have a smaller effect in larger populations. furthermore, we quantify the variance of the number of ancestral segments in a chromosome, and show that it is very large.

Similarly, in Chapter 3 we demonstrate that population subdivision only affects the expected number of ancestral segments in a substantial way if the population is very small. We also study segments in some regular mating systems, and present some original results for double-first-cousin and first-cousin mating systems.

Chapter 4 describes some models for the length of a tract IBD between two chromosomes, using some of the results of chapters 2 and 3. The best fitting model is one that allows for 1st order Markov dependence of junction types along the chromosome, and different average lengths for IBD and non-IBD segments. Application of this model to populations with different growth patterns and levels of subdivision again demonstrates that these factors are only important in small populations, and that the variance of the length of an IBD tract is large.

In Chapter 5, we describe some simulations of chromosome transmissions on the Hutterite pedigree and in random mating populations. The simulation programs themselves are a significant contribution, since they allow the realization of continuous chromosome data in reasonably large populations and pedigrees. We compare the results of simulation studies

to theoretical expectations that account for growth and large scale subdivision of the Hutterite population. The results suggest that there may be aspects of within-leut population structure that decrease the number of segments in a chromosome. However, the simulated values are not significantly different from those expected for a random mating population of the same historical sizes.

Finally, in Chapter 6, we develop a method of estimating small map distances using chromosomes sampled from the Hutterite population, and the pedigree relating sampled chromosomes. We use a quasi-likelihood model, and we find that modelling the covariance in the number of junctions in different chromosomes is required in order to get accurate variance estimates. We examine the performance of the method by simulation, and we find that there is substantial information within the Hutterite pedigree for estimating small map distances.

7.1 Implications for studies of disequilibrium in isolated populations

Our results show that the most important factor in determining the expected number of junctions in a chromosome, and therefore a lower bound for their average length, is the time since founding of the population. In generation t of an infinitely large random mating population, we expect t junctions per Morgan in a chromosome. In finite populations, the expectation is less than t, but the difference is substantial only if the historical population sizes have been small enough to result in the accumulation of IBD, and therefore the production of fewer junctions. Even when this is the case, the variance of the number of junctions in a chromosome is large, and so the existing number of junctions in a chromosome may differ substantially from that expected based on known population history and structure.

In light of these results, the findings of Lonjou et al. [22] are not surprising. They observed that for two particular small regions of the genome, samples from genetic isolates did not generally show stronger disequilibrium than samples from larger outbred populations. They did note strong disequilibrium in one isolate, the Ainu of Japan. Since only two regions of the genome were studied, there is a sample size of two for each population. Because variances are so large, little can be concluded from this. Also, they measured the

strength of disequilibrium in two small regions. Their paper therefore does not address how the *extent* of disequilibrium may differ between isolates and large populations.

Our theoretical results allow us to predict that disequilibrium might persist over longer distances in smaller, more recently founded populations. Whether or not it does depends on the patterns of junction formation in many meioses, which we cannot observe. Pilot studies of the extent of disequilibrium across the genome of an isolated population are therefore desirable, before the expense of a large scale disease mapping study is incurred.

Our approach to understanding disequilibrium in isolated populations is fundamentally different from traditional consideration of pairwise disequilibrium. We describe modern chromosomes as being composed of continuous segments of founder chromosomes – a mosaic of ancestral haplotypes. This viewpoint is particularly appropriate in the context of modern sequence data. For example, Maynard Olson reports (personal communication) that in a sample of Caucasians, all observed HLA haplotypes are the result of a small number of rearrangements of only two ancestral haplotypes.

7.2 Implications for estimation of genetic distance in isolated populations

We demonstrated in Chapter 6 that there is ample information in Hutterite chromosomes for the estimation of small map distances. The model presented required that for each sampled chromosome, we be able to count the number of junctions existing in the interval of interest. Since this is not possible, we considered the problem of detection and resolution of junctions from continous IBD data on all pairs of chromosomes in the sample. We found that for a random sample of 100 Hutterites, we could detect approximately 94% of existing junctions, and resolve about 75% of detected junctions. With further work we hope to be able to resolve probabilistically those junctions which are detected but not resolved, and impute the existence of those which are not detected directly. This requires a theoretical understanding of the relationship between the number of junctions detected and the number actually existing in the sampled chromosome, and application of an EM-type algorithm [14] to the quasi-likelihood model. For real data, it would be necessary to extend the method to apply to experimentally obtained haplotype data consisting of closely spaced markers

whose order along the chromosome is known.

The Hutterite population is an exciting and unique resource for the generation of fine-scale genetic maps. The age of the population, its isolation from other populations, and the availability of a complete and accurate pedigree make it ideal for map estimation using the approach we described. The typing of a large number of micro-satellite markers on a sample of several hundred Hutterites would provide a fine-scale map with much better precision than one obtained through traditional analysis of the CEPH pedigrees.

BIBLIOGRAPHY

[1] J. H. Bennett. Junctions in inbreeding. *Genetica*, 26:392–406, 1953.

[2] J. H. Bennett. The distribution of heterogeneity upon inbreeding. *Journal of the Royal Statistical Society*, 16:88–99, 1954.

[3] H. Blossey. *The Poisson Clumping Heuristic and the Survival of Genome in Small Pedigrees.* PhD thesis, University of Washington, Seattle, WA, 1993.

[4] N. H. Chapman and E. A. Thompson. Linkage disequilibrium mapping: The role of population history, size and structure. In *Advances in Genetics*, volume 42, pages 413–437. Academic Press Ltd, 2001.

[5] N. H. Chapman and E. M. Wijsman. Genome screens using linkage disequilibrium tests: Optional marker characteristics and feasibility. *American Journal of Human Genetics*, 63:1872–1885, 1998.

[6] J. F. Crow and M. Kimura. *An Introduction to Population Genetics Theory.* Harper and Row, New York, 1970.

[7] E. W. Daw, E. A. Thompson, and E. M. Wijsman. Bias in multipoint linkage analysis arising from map misspecification. *Genetic Epidemiology*, 19:366–380, 2000.

[8] J. W. Eaton and A. J. Mayer. *Man's Capacity to Reproduce.* The Free Press, Glencoe, Illinois, 1954.

[9] R. A. Fisher. *The Theory of Inbreeding.* Oliver and Boyd, Edinburgh, 1949.

[10] R. A. Fisher. A fuller theory of junctions in inbreeding. *Heredity*, 8:187–197, 1954.

[11] R. A. Fisher. An algebraically exact examination of junction formation and transmission in parent-offspring inbreeding. *Heredity*, 13:179–186, 1959.

[12] I. R. Franklin. The distribution of the proportion of the genome which is homozygous by descent in inbred individuals. *Theoretical Population Biology*, 11:60–80, 1977.

[13] J. C. Gale. Some applications of the theory of junctions. *Biometrics*, 20:85–117, 1964.

[14] C. C. Heyde and R. Morton. Quasi-likelihood and generalizing the EM algorithm. *Journal of the Royal Statistical Society, Series B*, 58:317–327, 1996.

[15] J. A. Hostetler. *Hutterite Society.* The Johns Hopkins University Press, Baltimore, 1974.

[16] J. A. Hostetler. *The Hutterites in North America.* Harcourt Brace College Publishers, Fort Worth, 1996.

[17] A. Jacquard. *The Genetic Structure of Populations.* Springer-Verlag, New York, 1974.

[18] S. Karlin and H. M. Taylor. *A First Course in Stochastic Processes.* Academic Press, San Diego, 1975.

[19] K. Knopp. *Infinite sequences and series.* Dover Publications Inc., New York, 1956.

[20] L. Kruglyak. Genetic isolates: Separate but equal? *Proceedings of the National Academy of Sciences*, 96:1170–1172, 1999.

[21] L. Kruglyak. Prospects for whole-genome linkage disequilibrium mapping of common disease genes. *Nature Genetics*, 22:139–144, 1999.

[22] C. Lonjou, A. Collins, and N. E. Morton. Allelic association between marker loci. *Proceedings of the National Academy of Sciences*, 96:1621–1626, 1999.

[23] P. McCullagh and J. A. Nelder. *Generalized Linear Models.* Chapman and Hall, New York, 1989.

[24] T. Nagylaki. Analysis of some regular systems of inbreeding. *Journal of Mathematical Biology*, 9:237–244, 1980.

[25] H. R. Nevanlinna. The Finnish population structure - a genetic and genealogical study. *Hereditas*, 71:195–236, 1972.

[26] A. Robertson. Artificial selection with a large number of linked loci. In *Proceedings of the International Conference on Quantitative Genetics*, pages 307–322. The Iowa State University Press, 1976.

[27] P. Stam. The distribution of the fraction of the genome identical by descent in finite random mating populations. *Genet. Res. Camb.*, 35:131–155, 1980.

[28] B. S. Weir, P. J. Avery, and W. G. Hill. Effect of mating structure on variation in inbreeding. *Theoretical Population Biology*, 18:396–429, 1980.

[29] A. Yu, C. Zhao, Y. Fan, W. Jang, A. J. Mungall, P. Deloukas, A. Olsen, N. A. Doggett, N. Ghebranious, K. W. Broman, and J. L. Weber. Comparison of human genetic and sequence-based physical maps. *Nature*, 409:951–953, 2001.

Appendix A

MOMENTS OF NON-IDENTITY PROPORTIONS IN A RANDOM MATING POPULATION

In this section, we calculate some moments of the non-IBD proportion in a random mating population, as defined at the beginning of Chapter 2. These are required for the development of expressions for $\mathrm{E}[n_i]$, $\mathrm{E}[n_i^2]$, and $\mathrm{E}[n_i n_j]$ in such a population, which we do in Appendix B.

Let $H_t(p)$ denote the proportion of the chromosome which is non-IBD in individual p of generation t. We can write

$$H_t(p) = \int_0^L \frac{I_t^p(x)}{L} dx \tag{A.1}$$

where L denotes the length of the chromosome in Morgans, and $I_t^p(x) = 1$ if individual p is non-IBD at point x on the chromosome, $I_t^p(x) = 0$ otherwise. Thus

$$\begin{aligned} \mathrm{E}[H_t(p)] &= \int_0^L \frac{\mathrm{E}[I_t^p(x)]}{L} dx \\ &= \int_0^L \frac{h_t(\underline{N})}{L} dx \\ &= h_t(\underline{N}) \ , \end{aligned} \tag{A.2}$$

where $h_t(\underline{N})$ is the probability of non-IBD at a particular locus in an individual in generation t.

In order to calculate higher order moments, we must consider some two-locus gene non-identity measures, described by Weir et al. [28], and illustrated in Figure A.1. Generally, we are interested in the probability that genes a and a' at locus x are non-IBD, and genes b and b' at locus y are also non-IBD. This probability is denoted Θ, Γ or Δ according to the number of chromosomes in which the loci are being compared (see Figure A.1). Weir et al. [28] consider the evolution of these probabilities over time, for populations reproducing according to various schemes of random mating with discrete generations. Let

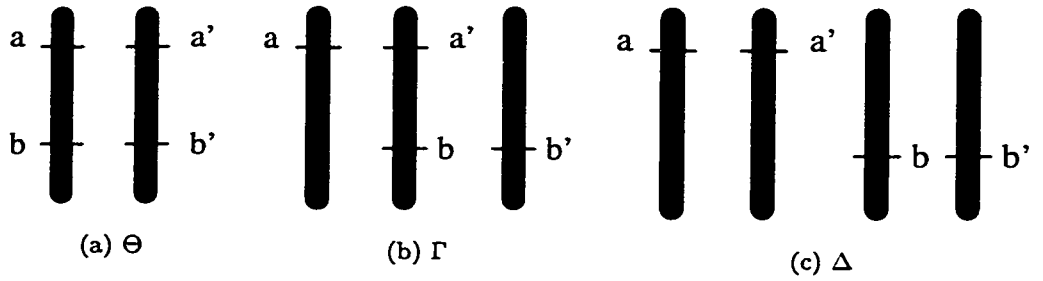


Figure A.1: Two locus gene non-identity measures.

$\nu_t = (\Theta_t, \Gamma_t, \Delta_t)'$ denote the two-locus non-IBD probabilities at generation t. Weir et al. [28] show that $\nu_{t+1} = \Omega \cdot \nu_t$, where Ω is a transition matrix which depends on the recombination fraction θ between the loci, and the size(N_t) of the population at generation t. Therefore v_t depends on the population sizes up to and including generation $t-1$, and the recombination fraction θ. We denote the probabilities of interest by $\Theta_t(\underline{N}, \theta)$, $\Gamma_t(\underline{N}, \theta)$, and $\Delta_t(\underline{N}, \theta)$.

We now consider $\mathrm{E}[H_t(p)^2]$, the expected value of the square of the non-IBD proportion in an individual in generation t.

$$\begin{aligned}
\mathrm{E}[H_t(p)^2] &= \mathrm{E}\left[\int_0^L \frac{I_t^p(x)}{L}\,dx \cdot \int_0^L \frac{I_t^p(y)}{L}\,dy\right] \\
&= \frac{1}{L^2}\int_0^L \int_0^L \mathrm{E}[I_t^p(x) \cdot I_t^p(y)]\,dx\,dy \\
&= \frac{1}{L^2}\int_0^L \int_0^L \Theta_t(\underline{N}, \theta_{|x-y|})\,dx\,dy \\
&= \frac{2}{L^2}\int_0^L (L-s)\Theta_t(\underline{N}, \theta_s)\,ds \\
&\equiv \overline{\Theta}_t(\underline{N}). \qquad \text{(A.3)}
\end{aligned}$$

In line 4 above, θ_s denotes the recombination fraction between two loci a distance s Morgans apart. Line 4 is obtained from line 3 by a change of variables. $\overline{\Theta}_t(\underline{N})$ is too complicated to evaluate exactly. Weir et al. [28] discuss its estimation by numerical integration.

We must also consider the product of the non-IBD proportions of two distinct individuals in the tth generation.

$$\mathrm{E}[H_t(p) \cdot H_t(p')] = \mathrm{E}\left[\int_0^L \frac{I_t^p(x)}{L}\,dx \cdot \int_0^L \frac{I_t^{p'}(y)}{L}\,dy\right]$$

$$
\begin{aligned}
&= \frac{1}{L^2}\int_0^L\int_0^L \mathrm{E}[I_t^p(x)\cdot I_t^{p'}(y)]\,dx\,dy \\
&= \frac{1}{L^2}\int_0^L\int_0^L \Delta_t(\underline{N},\theta_{|x-y|})\,dx\,dy \\
&= \frac{2}{L^2}\int_0^L (L-s)\Delta_t(\underline{N},\theta_s)\,ds \\
&\equiv \overline{\Delta}_t(\underline{N}). \qquad\qquad (A.4)
\end{aligned}
$$

Finally, we examine the product of the non-IBD proportions of two individuals, one from the ith generation, and one from the jth generation. We assume that $i<j$. Then

$$
\begin{aligned}
\mathrm{E}[H_i(p)\cdot H_j(p')] &= \mathrm{E}\left[\int_0^L \frac{I_i^p(x)}{L}\,dx\cdot\int_0^L \frac{I_j^{p'}(y)}{L}\,dy\right] \\
&= \frac{1}{L^2}\int_0^L\int_0^L \mathrm{E}[I_i^p(x)\cdot I_j^{p'}(y)]\,dx\,dy \\
&= \frac{1}{L^2}\int_0^L\int_0^L \Pr(a_i\neq a_i';\ b_j\neq b_j')\,dx\,dy\ , \qquad (A.5)
\end{aligned}
$$

where a_i and a_i' denote the genes at locus x in person p of generation i, b_j and b_j' denote the genes at locus y in person p' of generation j, and $\neq$ indicates non-IBD. In order to have $b_j\neq b_j'$, b_j and b_j' must be descended from different individuals in generation $j-1$. This implies that

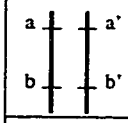
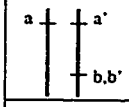
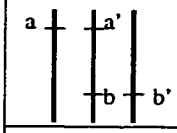
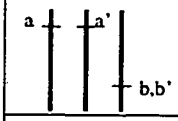
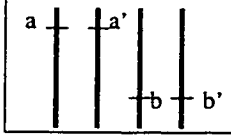
$$
\Pr(a_i\neq a_i';\ b_j\neq b_j') = \left(1-\frac{1}{2N_{j-1}}\right)\cdot\Pr(a_i\neq a_i';\ b_{j-1}\neq b_{j-1}'), \qquad (A.6)
$$

where b_{j-1} and b_{j-1}' denote the ancestors at generation $j-1$ of b_j and b_j' respectively. Applying A.6 iteratively, we obtain

$$
\Pr(a_i\neq a_i';\ b_j\neq b_j') = \prod_{k=1}^{j-i-1}\left(1-\frac{1}{2N_{i+k}}\right)\cdot\Pr(a_i\neq a_i';\ b_{i+1}\neq b_{i+1}'), \qquad (A.7)
$$

where b and b' denote genes at locus y on distinct chromosomes in generation $i+1$. The probability on the right hand side of Equation A.7 depends on the relationship between the chromosomes carrying a_i, a_i', and the ancestors b_i and b_i' of b_{i+1} and b_{i+1}'. Table A.1 shows the possible configurations of b_i and b_i', the probability of each configuration, calculated using the random mating model, and the desired probability $\Pr(a_i\neq a_i';b_{i+1}\neq b_{i+1}')$ conditional on that configuration.

Table A.1: Possible configurations of a, a', b, and b'.

Configuration	Probability	$\Pr(a \neq a'; b \neq b')$
	$2 \cdot \frac{1}{2N_i} \cdot \frac{1}{2N_i}$	$\Theta_i(\underline{N}, \theta)$
	$2 \cdot \frac{1}{2N_i} \cdot \frac{1}{2N_i}$	0
	$2 \cdot \frac{1}{2N_i} \cdot \frac{2N_i-2}{2N_i} \cdot 2$	$\Gamma_i(\underline{N}, \theta)$
	$\frac{2N_i-2}{2N_i} \cdot \frac{1}{2N_i}$	0
	$\frac{2N_i-2}{2N_i} \cdot \frac{2N_i-3}{2N_i}$	$\Delta_i(\underline{N}, \theta)$

The probability required in Equation A.5 is then obtained by summing over the possible configurations and substituting that quantity into Equation A.7. Therefore

$$\begin{aligned}\Pr(a_i \neq a_i';\ b_j \neq b_j') &= \prod_{k=1}^{j-i-1}\left(1-\frac{1}{2N_{i+k}}\right)\left[\frac{1}{2N_i^2}\Theta_i(\underline{N},\theta_{|x-y|})+\frac{2(N_i-1)}{N_i^2}\Gamma_i(\underline{N},\theta_{|x-y|})\right.\\ &\quad\left.+\frac{(N_i-1)(2N_i-3)}{2N_i^2}\Delta_i(\underline{N},\theta_{|x-y|})\right]. \qquad \text{(A.8)}\end{aligned}$$

Substituting Equation A.8 into Equation A.5, we find

$$\begin{aligned}\mathrm{E}[H_i(p)\cdot H_j(p')] &= \frac{1}{L^2}\int_0^L\int_0^L\prod_{k=1}^{j-i-1}\left(1-\frac{1}{2N_{i+k}}\right)\left[\frac{1}{2N_i^2}\,\Theta_i(\underline{N},\theta_{|x-y|})\right.\\ &\quad\left.+\frac{2(N_i-1)}{N_i^2}\,\Gamma_i(\underline{N},\theta_{|x-y|})+\frac{(N_i-1)(2N_i-3)}{2N_i^2}\Delta_i(\underline{N},\theta_{|x-y|})\right]\,dx\,dy\\ &= \prod_{k=1}^{j-i-1}\left(1-\frac{1}{2N_{i+k}}\right)\left[\frac{1}{2N_i^2}\,\frac{2}{L^2}\int_0^L(L-s)\Theta_i(\underline{N},\theta_s)\,ds\right.\\ &\quad+\frac{2(N_i-1)}{N_i^2}\,\frac{2}{L^2}\int_0^L(L-s)\Gamma_i(\underline{N},\theta_s)\,ds\\ &\quad\left.+\frac{(N_i-1)(2N_i-3)}{2N_i^2}\,\frac{2}{L^2}\int_0^L(L-s)\Delta_t(\underline{N},\theta_s)\,ds\right]\\ &= \prod_{k=1}^{j-i-1}\left(1-\frac{1}{2N_{i+k}}\right)\cdot\\ &\quad\left[\frac{1}{2N_i^2}\overline{\Theta}_i(\underline{N})+\frac{2(N_i-1)}{N_i^2}\overline{\Gamma}_i(\underline{N})+\frac{(N_i-1)(2N_i-3)}{2N_i^2}\overline{\Delta}_i(\underline{N})\right]. \qquad \text{(A.9)}\end{aligned}$$

Appendix B

MOMENTS OF THE NUMBER OF JUNCTIONS FORMED IN MEIOSES FROM A GIVEN GENERATION OF A RANDOM MATING POPULATION

To calculate $\mathrm{E}[n_i]$, $\mathrm{E}[n_i^2]$, and $\mathrm{E}[n_i n_j]$, we use the formula $n_i = \sum_{m=1}^{2N_{i+1}} X_i(m)$, where $X_i(m)$ denotes the number of junctions formed in meiosis m from generation i.

Since recombinations happen along the chromosome according to a Poisson process with rate one per Morgan,

$$X_i(m)|H_i(p_m) \sim Poisson(H_i(p_m)L), \tag{B.1}$$

where p_m denotes the parent of meiosis m, and L denotes the length of the chromosome in Morgans. Therefore

$$\begin{aligned} \mathrm{E}[X_i(m)] &= \mathrm{E}_H\left[\, \mathrm{E}[X_i(m)|H_i(p_m)] \,\right] \\ &= \mathrm{E}_H\left[H_i(p_m)L\right] \\ &= h_j(\underline{N})L\ , \end{aligned} \tag{B.2}$$

and

$$\begin{aligned} \mathrm{E}\left[X_i(m)^2\right] &= \mathrm{E}_H\left[\, \mathrm{E}[X_i(m)^2|H_i(p_m)] \,\right] \\ &= \mathrm{E}_H\left[H_i(p_m)L + (H_i(p_m)L)^2\right] \\ &= \mathrm{E}_H\left[H_i(p_m)]L + \mathrm{E}_H[H_i(p_m)^2\right] L^2 \\ &= h_i(\underline{N})L + \overline{\Theta}_i(\underline{N})L^2\ , \end{aligned} \tag{B.3}$$

since the parent is simply a randomly chosen individual from generation i. Then

$$\mathrm{E}[n_i] = \mathrm{E}\left[\sum_{m=1}^{2N_{i+1}} X_i(m)\right] = 2N_{i+1}h_i(\underline{N})L\ . \tag{B.4}$$

In order to calculate $\mathrm{E}[n_i^2]$, we also consider $\mathrm{E}[X_i(m)X_i(m')]$:

$$\begin{aligned}\mathrm{E}\left[X_i(m)X_i(m')\right] &= \mathrm{E}_H\left[\mathrm{E}[X_i(m)X_i(m')|H_i(p_m)H_i(p_{m'})]\right] \\ &= \mathrm{E}_H\left[H_i(p_m)H_i(p_{m'})L^2\right] \end{aligned} \tag{B.5}$$

since conditional on the proportion non-IBD in each of the parents, the numbers of junctions formed in each meiosis are independent. With probability $\frac{1}{N_i}$, both meioses are from the same parent. Otherwise they are from distinct individuals in the ith generation. Therefore

$$\begin{aligned}\mathrm{E}\left[X_i(m)X_i(m')\right] &= \mathrm{E}_H\left[H_i(p_m)H_i(p_{m'})L^2\right] \\ &= \frac{1}{N_i}\mathrm{E}_H\left[H_i(p)^2L^2\right] + \frac{N_i-1}{N_i}\mathrm{E}_H\left[H_i(p)H_i(p')L^2\right] \\ &= \frac{1}{N_i}\overline{\Theta}_i(\underline{N})L^2 + \frac{N_i-1}{N_i}\overline{\Delta}_i(\underline{N})L^2. \end{aligned} \tag{B.6}$$

Then

$$\begin{aligned}\mathrm{E}[n_i^2] &= \mathrm{E}\left[\left(\sum_{m=1}^{2N_{i+1}} X_i(m)\right)^2\right] \\ &= \mathrm{E}\left[\sum_{m=1}^{2N_{i+1}} X_i(m)^2\right] + \mathrm{E}\left[\sum_{m=1}^{2N_{i+1}}\sum_{\substack{m'=1,\\ m'\neq m}}^{2N_{i+1}} X_i(m)X_i(m')\right] \\ &= 2N_{i+1}\left(h_i(\underline{N})L + \overline{\Theta}_i(\underline{N})L^2\right) + \\ &\quad 2N_{i+1}\cdot(2N_{i+1}-1)\left(\frac{1}{N_i}\overline{\Theta}_i(\underline{N}) + \frac{N_i-1}{N_i}\overline{\Delta}_i(\underline{N})\right)L^2. \end{aligned} \tag{B.7}$$

Calculation of $\mathrm{E}[n_in_j]$ requires that we first obtain $\mathrm{E}[X_i(m)X_j(m')]$, the expected value of the product of the numbers of junctions formed in two meioses occurring in two different generations.

$$\begin{aligned}\mathrm{E}\left[X_i(m)X_j(m')\right] &= \mathrm{E}_H\left[\mathrm{E}[X_i(m)X_j(m')|H_i(p_m)H_j(p_{m'})]\right] \\ &= \mathrm{E}_H\left[H_i(p_m)H_j(p_{m'})L^2\right] \\ &= \prod_{k=1}^{j-i-1}\left(1-\frac{1}{2N_{i+k}}\right)\cdot\left[\frac{1}{2N_i^2}\overline{\Theta}_i(\underline{N})\right. \\ &\quad \left.+\frac{2(N_i-1)}{N_i^2}\overline{\Gamma}_i(\underline{N}) + \frac{(N_i-1)(2N_i-3)}{2N_i^2}\overline{\Delta}_i(\underline{N})\right]L^2. \end{aligned} \tag{B.8}$$

for $i < j$, by Equation A.9. Then

$$\begin{aligned}
\mathrm{E}[n_i n_j] &= \mathrm{E}\left[\sum_{m=1}^{2N_{i+1}} X_i(m) \cdot \sum_{m'=1}^{2N_{j+1}} X_j(m')\right] \\
&= \mathrm{E}\left[\sum_{m=1}^{2N_{i+1}} \sum_{m'=1}^{2N_{j+1}} X_i(m) X_j(m')\right] \\
&= \sum_{m=1}^{2N_{i+1}} \sum_{m'=1}^{2N_{j+1}} \mathrm{E}\left[X_i(m) X_j(m')\right] \\
&= 2N_{i+1} 2N_{j+1} \prod_{k=1}^{j-i-1} \left(1 - \frac{1}{2N_{i+k}}\right) \cdot \left[\frac{1}{2N_i^2} \overline{\Theta}_i(\underline{N})\right. \\
&\quad \left. + \frac{2(N_i - 1)}{N_i^2} \overline{\Gamma}_i(\underline{N}) + \frac{(N_i - 1)(2N_i - 3)}{2N_i^2} \overline{\Delta}_i(\underline{N})\right] L^2. \qquad \text{(B.9)}
\end{aligned}$$

VITA

Nicola H. Chapman was born on June 24th, 1971, in Guelph, Ontario, Canada. She earned a Bachelor of Science degree in Statistics from the University of Toronto in 1993, and a Master of Science degree in Biostatistics, also from the University of Toronto, in 1995. She moved to Kirkland, Washington in 1995, and earned a Doctor of Philosophy in Biostatistics from the University of Washington in 2001. She currently lives in Kirkland, Washington with her husband, and a really great dog.