Department of Computer Science
Series of Publications A
Report A-2002-2

Simulation Techniques for Disease Gene Localization in Isolated Populations

Vesa Ollikainen

University of Helsinki
Finland

Department of Computer Science
Series of Publications A
Report A-2002-2

Simulation Techniques for Disease Gene Localization in Isolated Populations

Vesa Ollikainen

To be presented, with the permission of the Faculty of Science of the University of Helsinki, for public criticism in Auditorium XII, University Main Building, on August 16th, 2002, at 12 o'clock.

University of Helsinki
Finland

Contact information

Postal address:
Department of Computer Science
P.O.Box 26 (Teollisuuskatu 23)
FIN-00014 University of Helsinki
Finland

Email address: Vesa.Ollikainen@cs.Helsinki.FI

URL: http://www.cs.Helsinki.FI/

Telephone: +358 9 1911

Telefax: +358 9 191 44441

ISSN 1238-8645
ISBN 952-10-0614-5
Computing Reviews (1998) Classification: G.3, I.6.3, J.3
Helsinki 2002
Helsinki University Printing House

Simulation Techniques for Disease Gene Localization in Isolated Populations

Vesa Ollikainen

Department of Computer Science
P.O. Box 26, FIN-00014 University of Helsinki, Finland
Vesa.Ollikainen@cs.helsinki.fi, http://www.cs.helsinki.fi/u/vollikai/

PhD Thesis, Series of Publications A, Report A-2002-2
Helsinki, August 2002, 149+5 pages
ISSN 1238-8645, ISBN 952-10-0614-5

Abstract

In this thesis we present new simulation techniques and algorithms for localizing disease genes for human multifactorial diseases, which are disorders controlled by several genetic, environmental, and stochastic risk factors. From a computational point of view, gene localization is a complex problem.

We introduce a population simulator software package for exploring the complex relationships between the models of multifactorial disease and localization outcomes. The population simulator consists of components for simulating pedigree generation according to the population history specification given by the user, the inheritance of chromosomal segments in the pedigrees, the emergence of complex disease phenotypes, and the collection of samples. The simulator can be used to estimate the significance of a positive localization outcome, to estimate the power of the proposed study setup, to analyze the connection between disease models and population history, and to perform comparative studies on different methods and measures used in disease gene localization.

We concentrate on haplotype association methods that take advantage of linkage disequilibrium (nonrandom association of alleles in closeby loci) and describe a method for using the population simulator in analyzing the effect of different characteristics of population history (such as population age, size, and substructure) on the amount of linkage disequilibrium present in the data.

We also describe a computational framework and a software implementation for finding the optimal gene mapping strategy for a genome-wide search for genes of complex diseases in an isolated population. The procedure is applied for different population histories, disease models, and measures of linkage disequilibrium.

We combine the developed simulation framework with three established gene localization methods (Haplotype relative risk, Haplotype pattern mining and Transmission/disequilibrium tests), and analyze their applicability in the initial stage of a genomewide search.

Computing Reviews (1998) Categories and Subject Descriptors:
G.3 [Probability and Statistics]: Nonparametric statistics, Probabilistic algorithms
I.6.3 [Simulation and Modeling]: Applications
J.3 [Life and Medical Sciences]: Biology and genetics

General Terms:
Algorithms, Design, Experimentation, Measurement

Additional Key Words and Phrases:
Population simulation, Gene mapping, Linkage disequilibrium, Haplotype analysis

Acknowledgements

I am extremely grateful to my supervisor Professor Heikki Mannila for his invaluable guidance and support during the preparation of this work. Especially, I want to thank him for his warm and encouraging words that made me continue with the thesis and rediscover the enthusiasm at the moment when my motivation was at stake.

I want to thank the current and former heads of the three departments I have been working at: I am thankful to Professor Juha Kere, who created an inspiring atmosphere in the Finnish Genome Center. I would like to thank Professor Elja Arjas for an interesting year at Rolf Nevanlinna Institute, which brought me in touch with new perspectives. I appreciate that Department of Computer Science, currently headed by Professor Jukka Paakki, has always provided me with excellent working conditions.

I was privileged to work with Ph.D. Päivi Onkamo, M.Sc. Petteri Sevon, M.Sc. Kari Vasko, and Professor Hannu Toivonen on several exciting topics and ideas. I really enjoyed the projects and the numerous discussions we had.

I am grateful to Professor Martti Juhola and Ph.D. Mikko Sillanpää for their invaluable comments. I also highly appreciate the cooperation and discussions, both formal and informal, that we had with M.Sc. Hanni Kärkkäinen, M.Sc. Riika Kilpikari, M.Sc. Sampo Smolander, M.Sc. Mikko Koivisto, Ph.D. Pekka Uimari, and M.D., Ph.D. Tarja Laitinen.

I would like to thank M.A. Marina Kurtén for correcting the grammar of this thesis. Also, I want to thank all the unmentioned workmates at Rolf Nevanlinna Institute, Finnish Genome Center and Department of Computer Science, including HIIT Basic Reasearch Unit, who it has been a pleasure to work with.

I am thankful to the Graduate School of Computational Biology, Bioinformatics, and Biometry for making this work financially possible. I especially appreciate the ComBi courses and meetings that provided valuable contacts as well as interesting insights into related research areas.

I am most indebted to my parents Liisa and Eero as well as my grandparents Hilkka and Martti who have constantly supported me during these years. Finally, I want to express by sincerest gratitude to my wife Elina for her love, patience, and understanding.

Helsinki, July 23, 2002
Vesa Ollikainen

Contents

1 Introduction **1**

2 Preliminaries **5**
2.1 Fundamentals of human genes and chromosomes 5
2.2 Markers and genetic data 6
2.3 Sharing identical-by-descent and identical-by-state 10
2.4 Genetic linkage 11
2.5 Linkage disequilibrium and Hardy-Weinberg equilibrium . . 12
2.6 Genetic disease and genetic models 14

3 Motivation of the thesis **17**

4 A population simulator **21**
4.1 Introduction 21
4.2 Simulating population history 24
4.2.1 Homogeneous populations, overlapping generations . 24
4.2.2 Substructured populations, distinct generations . . . 31
4.3 Simulating inheritance of chromosomes 32
4.4 Simulating disease phenotype and sample collection 42
4.5 Simulating generation of marker data 47
4.6 Simulation-based estimation of significance and power . . . 47

5 For maximal power at minimal cost **51**
5.1 Approaches to disease gene localization 51
5.1.1 Linkage analysis 52
5.1.2 Association and linkage disequilibrium analysis . . . 54
5.1.3 Combining linkage and association 63
5.1.4 Multiple testing problem and its solutions 64
5.2 Power and cost of disease gene localization 66
5.2.1 Stages, power, and false positives 67
5.2.2 Obtaining the power function 69

5.2.3 Cost components and minimization 71
5.2.4 Example of cost computation 73

6 Population history and conserved chromosomal segments 77
6.1 The experiment design 78
6.2 Results and conclusions 82

7 Detection of disease–marker association 91
7.1 Biallelic measures 91
7.2 Multiallelic measures 92
7.3 Populations and cost models 93
7.4 Single source, fixed prevalence 98
7.4.1 The simulation procedure 98
7.4.2 Disease models, marker maps, and sample sizes . . . 98
7.4.3 Computation of the power function 101
7.4.4 Results 102
7.5 Multiple sources, variable prevalence 103
7.5.1 The simulation characteristics 107
7.5.2 Disease model, marker maps, and sampling strategy 110
7.5.3 Results 111

8 Population history and disease gene localization 115
8.1 The experimental design 116
8.2 Effect of population history 118
8.2.1 Population substructure and migration 118
8.2.2 Size and shape of population pedigree 121
8.3 Effect of number of founder mutations 122
8.4 Effect of study design 122
8.4.1 Marker map 122
8.4.2 Method of linkage disequilibrium mapping 123
8.4.3 Effect of sampling strategy: family-based vs. population-based control chromosomes 124

9 Challenge of multifactorial diseases 129

10 Conclusions 137

References 141

Appendix A A–1

Chapter 1

Introduction

This thesis focuses on simulation techniques in finding disease genes for human multifactorial diseases. The emphasis is on methods that have applications in gene mapping in isolated populations, such as Finnish subpopulations. In this introductory chapter we take a look at how this topic fits into the fields of computer science, computational biology and medical genetics in general. A more detailed problem statement and contributions of this thesis will follow in Chapter 3.

Genetics-related topics have become popular in the mass media during the last few years, and one of the most remarkable breakthroughs has been the completion of the human DNA sequence [LLB01, VAM01]. While the topic is definitely titillating, among the most important consequences of the recent advances — and hopefully the one that turns out to be of great importance to mankind — is the rapidly increasing knowledge in medical genetics and its application in the prevention and cure of human diseases. Increased biomedical knowledge can lead to better understanding of the functions of the human body, deeper understanding of the mechanisms underlying common diseases, better diagnostic tests, and more effective medication with fewer adverse effects. However, the processes of turning biological data to knowledge and, finally, to efficient treatment of diseases, take time and require not only biological but also solid mathematical, statistical and computer science expertise. The new fields of computational biology and bioinformatics have emerged to answer these needs.

The publication of the human DNA reference sequence has allowed scientists to predict the locations of genes and draw conclusions about factors such as gene quantity and distributions of different bases in the sequence. However, the relationship between genes and diseases is, in most cases, still unknown. It is a popular metaphor to compare the DNA sequence to a book written in a foreign language: even though all the words are there, we still cannot understand the message. Currently, we are at a point where the meaning of some words is known, but we still have not found the rules, how the words are combined into sentences and paragraphs. Learning this new language, or investigating how the genes affect the development of the disease, is one of the major challenges in biomedical research and bioinformatics.

In large gene mapping projects, vast quantities of data from affected and unaffected individuals are collected, or at least available with a reasonable effort. As a consequence, the biomedical problem of finding the genes has turned into a plethora of interesting computational and statistical problems: how can we distinguish between the true and false signals; how to choose the best study strategy, and how should we measure the relationship between the genetic factors and the disease; how should the environmental factors be taken into account in the analysis, and how can we handle the possible interaction of multiple genes? Since it is not possible to analyze all chromosome locations where there is variation between individuals explicitly, at what resolution should we look at the genome? What do we understand about the effect of population history that in the end has had a major contribution of what our data looks like, and how are those special characteristics efficiently utilized in the gene localization process?

From a computer scientist's point of view, the challenge is in creating the sufficient computational means to facilitate answering some of these questions. These computational means include methods for testing different study setups, and algorithms for efficient gene localization, to name a few.

The concept of statistical gene mapping dates back to the 1950's [Mor55], when the idea of linkage analysis was introduced. For decades, the successes in gene mapping were, however, related to simple and relatively rare monogenic diseases. Now we are at a point where most diseases following a simple Mendelian inheritance pattern are revealed, and the focus has shifted towards multifactorial, or complex, diseases where there are multiple genes contributing to the disease, possibly in interaction with complex, unknown environmental factors. As an additional complication, the definition of the disease phenotype itself can be difficult.

In spite of the difficulties caused by the nature of multifactorial diseases, Finland has one advantage over many countries in gene mapping: its population history. The Finnish population is of relatively young age, and it contains isolated regions where immigration has been very modest. As a consequence of the isolation, it is expected that the genetic basis of the disease is more unified, which eventually makes the localization process easier [Del93, DeW98]. To take full advantage of that property is, again, not easy. The relationship between population history and observed data is not clear, and there is a demand for methods that help reveal the nature of this relationship. In this work we create a simulation framework that allows taking advantage of the underlying population history in the study design phase. Throughout the thesis we focus on an important subset of diseases, where the affection status is of qualitative nature, i.e., each individual can be classified to be either 'affected' or 'healthy'.

The thesis begins with a preliminaries chapter (Chapter 2), where a sufficient background on the concepts in disease gene localization is given. This lays the foundation for the presentations of the detailed research problems in Chapter 3. Next, the means for achieving the research tasks are presented in Chapter 4, where we describe the simulation methods and software that were designed and implemented as a part of this work, and used in the experiments of the subsequent chapters. Chapter 5 contains a literature review of widely used gene localization methods, followed by the presentation of our simulation-based framework for evaluating the performance of different gene mapping strategies and finding the optimal study strategy for a genome mapping project. The next chapter (Chapter 6) provides information about the limits of gene mapping in isolated populations; in that chapter we analyze, how much genetic material individuals living in isolated populations share with each other and with the ancestral founders that are thought to have brought the disease genes into the population isolate. This gives an idea of what the "background noise" in the data is expected to be, and what the overall resolution in looking at variable sites (or, in genetics terms, the density of the marker map) should be.

In Chapter 7, the task is to find suitable statistics for measuring the association of the disease to different alleles (i.e., variants in the DNA samples). This is an important issue, since the advent of new laboratory technologies has made it possible to extract information from multiple, closely located polymorphic sites with two variants (single nucleotide polymorphisms, or SNPs). From each individual, there is an increased number of data points available, and an increased probability to obtain false positive signals. The key issue is how to efficiently combine the information from multiple polymorphic sites to maintain the balance between the power of the study and the rate of false positive signals.

From the simulation study of different measures of linkage disequilibrium, we move further towards gene localization in reality. In Chapter 8, the performance of three established gene localization methods are compared with simulated data sets, and the effect of population history on disease gene localization power is analyzed for each one of them. This creates a direct link between the population history and the real, observed performance in gene localization as we know it today.

In Chapter 9 we consider the multiple-locus reality of common multifactorial diseases and seek the limits of the mapping methods that focus on one disease-predisposing gene at a time. Finally, Chapter 10 concludes this work.

Chapter 2

Preliminaries

2.1 Fundamentals of human genes and chromosomes

In this chapter we get acquainted with the key concepts required for the computational analysis of human genomic data. A description of the structure of the human genome can be found in any introductory textbook of human molecular genetics, such as [StR96], or [JCW95], and much of the information presented in this chapter is based on the contents of those books.

Human genetic data is stored in *chromosomes*, which are linear bodies containing DNA. A normal human genome consists of 46 chromosomes. 44 of these 46 chromomes are structurally pairwise similar, or *homologous*. These homologous chromosomes are referred to as *autosomes*. The autosomes form 22 *chromosome pairs*, which are generally referred to as human chromosomes 1 to 22. In addition to the 22 pairs of autosomes, the normal human genome consists of two *sex chromosomes*: a female has one pair of X chromosomes whereas a male has one X and one Y sex chromosome.

The genome of a human being, which thus consists of human chromosomes $1, \ldots, 22$ (two copies of each) and two sex chromosomes, is present in every cell of an individual. The only exception to this rule are the sperm and egg cells, where, because of a cell division process called *meiosis*, only one copy of each autosome and one sex chromosome is present.

Chromosomes, comprised of double-stranded DNA, contain *genes*, whose main function is to code for proteins. The genes are scattered over the genome among chromosomal segments of no information. The genetic knowledge is stored in the DNA in the form of base pairs. It is probable that only one tenth of the 3 billion base pairs in the human genome has a coding function. Until recently, human beings were believed to have 80,000 or more genes, but as a result of sequencing of human DNA, the estimate has dropped drastically to 30,000–40,000 genes [LLB01, VAM01].

The location of a large proportion of genes is known, but the function of the majority of the genes is yet unknown. Moreover, to locate the coding regions within the genes is not always trivial.

The position of a particular gene on a chromosome is called a *locus*. Each individual has two *alleles* at each locus of an autosomal chromosome pair, one in each homologous chromosome. An offspring normally inherits its two copies of each of the 22 autosomes, one from the mother and the other from the father. However, the copies inherited are not always exact replicates of the parents' chromosomes. In genesis of germline cells, the two homologous chromosomes in each chromosome pair may exchange parts of their DNA segments with each other, which results in new combinations of genetic material (see Figure 2.1). This phenomenon takes place in an early phase of gamete formation, *meiosis*, and is called *crossover*. The probability of two loci in the same chromosome being separated by a crossover is roughly proportional to the physical distance between the loci. The genetic length of a chromosome is measured in *centiMorgans* (abbreviated cM), one centiMorgan being equal to the length of a segment that is expected to be split by a crossover event once in a hundred meiotic steps. According to the Genethon human linkage map [DFF96], the sex-averaged genetic lengths of human chromosomes vary between 58 cM (chromosome 22) and 293 cM (chromosome 1). Due to the supposedly higher crossover frequency in females, female genetic lengths are assumed to exceed those of males.

2.2 Markers and genetic data

In search of disease-predisposing genes, the crossovers cannot be observed directly. Instead, the chromosomes are sampled at fixed loci, where variation exists between individuals. These loci are called *markers*. The two *alleles*, or variants, of an individual at a marker locus can be determined in a laboratory almost unequivocally. A marker is considered highly informative, if it has several alleles which are distributed in nearly equal proportions among the population. A *microsatellite* marker can have up to 20 different alleles, each of which corresponds to a different number of repetitions in the individual's DNA sequence. *SNP* (single nucleotide polymorphism) markers, on the other hand, nearly always have only two alleles.

For a given chromosome c and individual i, we can now define the (multilocus) *genotype* of individual i covering n markers as a vector of unordered

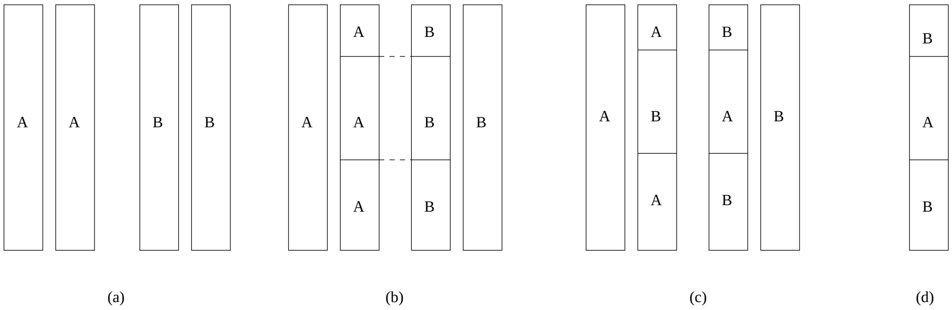

Figure 2.1: A schematic figure of a meiosis event in a pair of autosomes. Meiosis is a process in formation of germline cells, or *gametes*. The individual has inherited two copies of each autosome, one from the father (denoted by A) and one from the mother (denoted by B). (a) In the meiosis, the DNA in the two chromosomes is duplicated. (b) The duplicated chromosomal strands (called *chromatids*) exchange genetic material. In this example, two *chiasmata* are formed, denoted by the horizontal lines. (c) The two chromatids have exchanged their genetic material in the region between the two chiasmata. We say that *crossovers* have occurred at the locations of the chiasmata. (d) In female meiosis, the DNA in a random chromatid (out of the four chromatids altogether) will be included in the resulting gamete. In male meiosis, all four chromatids of figure (c) can end up in four different gametes.

allele pairs

$$G_{i,c} = (\{a_{11}, a_{12}\}, \ldots, \{a_{n1}, a_{n2}\}).$$

To return to Figure 2.1, we recall that the chromatid in Figure (d) may subsequently be passed to the offspring of the individual. Because of the recent crossover events (two in this case), the offspring will then have obtained genetic material from two grandparents (the segments labeled as A from one grandparent and those labeled as B from the other). The same process naturally happens in the gamete formation of the other parent, so an individual may have inherited chromosomal segments either from 2, 3, or 4 grandparents in each chromosome.

To illustrate how the past inheritance of chromosomal segments can be observed using genetic markers, consider the situation of Figure 2.2, which shows the chromosomal segments of a father, a mother and an offspring for the human chromosome c' (where $c' \in \{1, \ldots, 22\}$). In the example, three markers ($M1$, $M2$, and $M3$) are placed on the chromosome, and the genotype of each individual is observed in each marker locus. Applying the notation given above, we write the multilocus genotypes of the father (F), mother (M), and the offspring (O) as $G_{F,c'} = (\{1,2\},\{1,2\},\{2,3\})$, $G_{M,c'} = (\{3,4\},\{3,4\},\{1,2\})$, and $G_{O,c'} = (\{2,3\},\{1,4\},\{1,3\})$. Note that the genotypes in each marker locus are unordered pairs, since it is impossible to distinguish between paternally and maternally inherited alleles by laboratory methods.

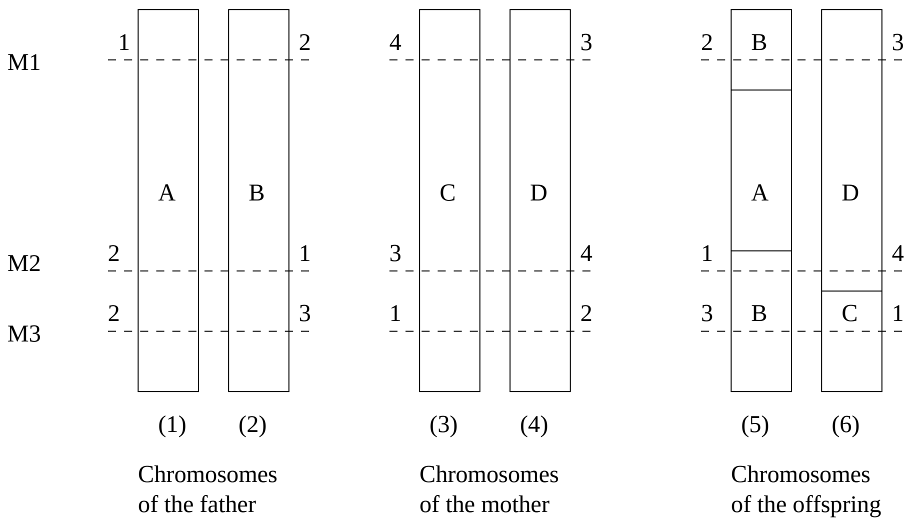

Figure 2.2: An example of inheritance of chromosomes and marker alleles from two parents to an offspring. Letters A to D indicate the original chromosomal segments of the two parents. Due to crossover events in the meiosis, the chromosome passed to the offspring from the father (5) contains genetic material from both of the paternal chromosomes (1) and (2), and, likewise, the chromosome passed to the offspring from the mother (6) contains genetic material from both of the chromosomes of the mother (3) and (4). The crossover sites of the paternal meiosis are indicated by the two solid horizontal lines in chromosome (5) and the crossover site of the maternal meiosis by the solid horizontal line in chromosome (6). The three marker loci ($M1$, $M2$, and $M3$) are indicated by dashed lines, and the integers at the marker loci correspond to marker alleles.

In cases where samples are also collected from several members of the pedigree, it may be possible to specify the parental origin of each allele. In this case, the individual's *haplotypes* can be written as:

$$H^1_{i,c} = (a_{11}, a_{21}, \ldots, a_{n1}), H^2_{i,c} = (a_{12}, a_{22}, \ldots, a_{n2}).$$

The haplotype representation couples each allele in the genotype with either of the two chromosomes in the chromosome pair of the individual.

Next, assume that the crossover sites and chromosomal segments of Figure 2.2 were hidden, and it was our task to deduce the six haplotypes based on genotypes $G_{F,c'}$, $G_{M,c'}$, and $G_{O,c'}$ only. By looking at the alleles shared by the offspring and each of the parents, we can in this case deduce that $H^1_{O,c'} = (2, 1, 3)$, and $H^2_{O,c'} = (3, 4, 1)$.

In addition, we can deduce that haplotype $H^1_{O,c'}$ of the offspring has been inherited from the father and haplotype $H^2_{O,c'}$ from the mother. However, we cannot reliably deduce the four parental haplotypes because of possible recombinations in the parental meioses.

It is important to observe that the deduction of the haplotypes using pedigrees can and must be based solely on genotype data ($G_{F,c'}$, $G_{M,c'}$, and $G_{O,c'}$), since we do not have any direct information about the crossover events of the parents. On the other hand, the construction of haplotypes from the genotype data allows us to make *some* deduction about the locations, where crossover events have taken place.

If the haplotypes of the two parents were known (based on genotypes of the grandparents, for instance), we could see that haplotype $H^2_{O,c'}$ is a combination of two haplotypes of the mother: the alleles in the first two marker loci have been inherited from haplotype $H^2_{M,c'}$ and the allele in the third marker locus from haplotype $H^1_{M,c'}$. In this case, we would say that a *recombination* has occurred between loci $M2$ and $M3$.

If we had the hidden crossover information of Figure 2.2 at our disposal, we could see that the observed recombination is a consequence of a crossover event between the two marker loci. Without that information, we could not, however, rule out the possibility that the recombination was caused by 3, 5, or generally $2n + 1$ ($n \in N$) crossover event between markers $M2$ and $M3$. Thus, there is not always a one-to-one correspondence between (observed) recombinations and (hidden) crossovers.

Likewise, we see that haplotypes $H^2_{F,c'}$ and $H^1_{O,c'}$ are identical. Based on the haplotype data, we do not thus observe any recombinations in the inheritance of the chromosome from the father to the offspring. By looking at the hidden information of Figure 2.2, however, we see that there have been two (unobserved) crossover events between markers $M1$ and $M2$. In general, any even ($2n$, $n \in N$) number of crossover events between those markers would have resulted in a similar observation.

These examples allows us to link the concept of (observed) recombinations to that of (hidden) crossover events in a meiosis: a recombination has occurred between markers M_i and M_j in the same chromosome if there has been an odd number of crossovers in the chromosome region between the two loci.

In the previous example, the presence of pedigree information and parental genotypes made it possible to deduce the haplotypes of the offspring. However, sometimes parental samples are not available, and the analysis must be based on genotype data, or, alternatively, the determination of haplotypes must be based on genotypes of (unrelated) individuals collected from the same population. Such haplotype construction by utilizing population data has been shown possible (see [SSD01] for example) in cases where densely spaced markers are available.

Recall that in the haplotype representation each allele in the marker locus was assigned to either of the two chromosomes of an individual. We call such haplotypes either *observed* (when there is no ambiguity in construction of the haplotypes) or *hypothetical* (haplotypes that are consistent with the observed genotypes, even though not necessarily 'correct'). Frequently, it is necessary to refer to parts of haplotypes that do not necessarily span all markers of the chromosome. For this purpose, a *subhaplotype* in a map of n markers is defined as a sequence of alleles of one or more adjacent marker loci, i.e., $(a_i, \ldots, a_j)$, where $1 \leq i \leq j \leq n$, and $a_i, \ldots, a_j$ are alleles of an observed or hypothetical haplotype of an individual. The term 'haplotype' is, however, often used as a synonym for subhaplotype, and we will also do so in the future sections.

The genotype and haplotype data, as described here, are used to find the approximate positions of trait-influencing genes. These methods will be discussed in the subsequent chapters.

2.3 Sharing identical-by-descent and identical-by-state

The use of genetic markers requires distinguishing between two concepts: *identical-by-state* (IBS) sharing and *identical-by-descent* (IBD) sharing. Two individuals are said to share an allele IBS, if the alleles of the individuals carry a common label, i.e., are indistinguishable by laboratory methods. Two individuals share an allele identical-by-descent, if they share it IBS, and, in addition, it is known that these individuals must have inherited an identical copy of the allele from the same ancestor. From the genotype data, IBS-sharing in a marker locus is easy to observe, but observation of

IBD-sharing requires deduction based on pedigree information, and may not always be possible.

The concepts of IBS- and IBD-sharing are illustrated in Figure 2.3. In the three *nuclear pedigrees* of two parents and their children, the IBS status can always be unambiguously determined. However, determination of the IBD status is not possible for pedigree (c).

The fact that IBS-sharing is not a proof of IBD-sharing is a problem only when the marker information content is low. In cases where the markers are highly polymorphic, IBS-sharing indicates IBD-sharing with high probability.

2.4 Genetic linkage

Two loci are inherited independently, if they reside in different chromosomes. Likewise, alleles in two loci wide apart in the same chromosome are inherited (nearly) independently. With shorter distances, there is an increased probability that there are no crossover events between the loci, and the alleles in the same chromatid are passed to the offspring unchanged. In this case, the loci are said to be *genetically linked.*

An example of genetic linkage is given in Figure 2.4. The map contains three markers, out of which marker A is located apart from the other two markers, B and C, which in turn are physically close to each other. The figure shows the inheritance of the genotypes to the offspring as well as the 'true' haplotype configurations. From the observed data we can in this case count that the alleles in loci A and B had been inherited together to an offspring 10 times, and a recombination had occurred 4 times[1]. From this data, we can compute an estimate for the *recombination fraction* between loci A and B to be $\theta_{AB} = 4/14 \approx 0.29$. Likewise, the recombination fraction between loci B and C can be approximated to be $\theta_{BC} = 1/14 \approx 0.07$, and between loci A and C to $\theta_{AC} = 4/14 \approx 0.29$. (This example illustrates that the observed recombination fractions do not generally behave in an additive way. For an additive measure of genetic length, one must convert the recombination fractions between adjacent markers to centiMorgans (see page 2.1). For this purpose, different *mapping functions* exist). The low recombination fractions between pairs of markers in the example indicate that the two loci are *linked.* The linkage between markers B and C is stronger than that between markers A and B. However, it should be noticed that

[1]Recall from the definitions of crossover and recombination that counts 10 and 4 correspond to cases where an even and odd number of crossovers have occurred, respectively. With small genetic lengths, the probability of more than one crossover is very low.

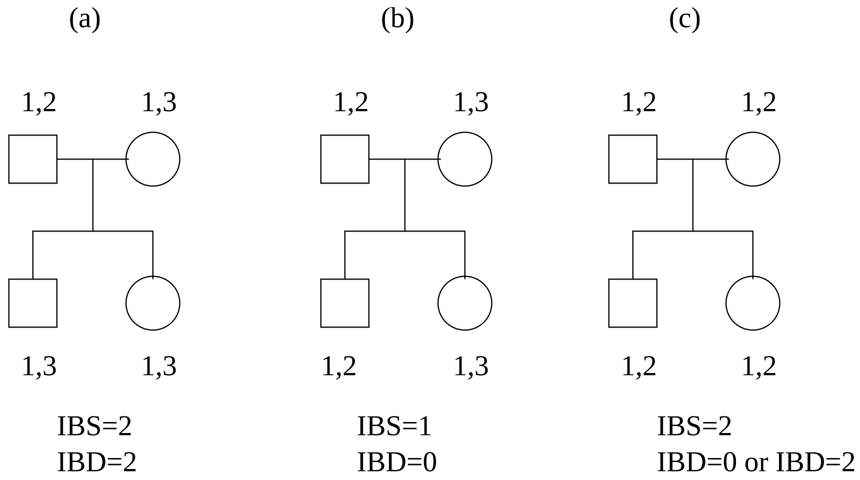

Figure 2.3: Three nuclear pedigrees of two parents and their children, the genotypes of the family members in a single marker locus, and the status of sharing identical-by-state (IBS) and identical-by-descent (IBD) between the two offspring in each pedigree. The symbols in the graphical representation of the pedigrees are by convention: square for male, circle for female. Spouses are connected by a horizontal line, and the vertical line downwards connects to the children of the couple. (a) The inheritance of the alleles can be unambiguously determined. The offspring share both alleles IBS and IBD. (b) The inheritance of the alleles can also be unambiguously determined. The offspring share allele 1 IBS, but, since they have inherited it from different parents, the offspring do not share any alleles IBD. (c) It is impossible to determine the inheritance of alleles. The offspring share 2 alleles IBS, but, assuming Mendelian inheritance, the IBD status can be with equal probability 0 or 2.

estimating the recombination fractions from two modestly-sized pedigrees like in this example is extremely inaccurate.

2.5 Linkage disequilibrium and Hardy-Weinberg equilibrium

Repeated crossovers eventually lead to a vast number of gene combinations, which explains the genetic variation between individuals. A chromosome

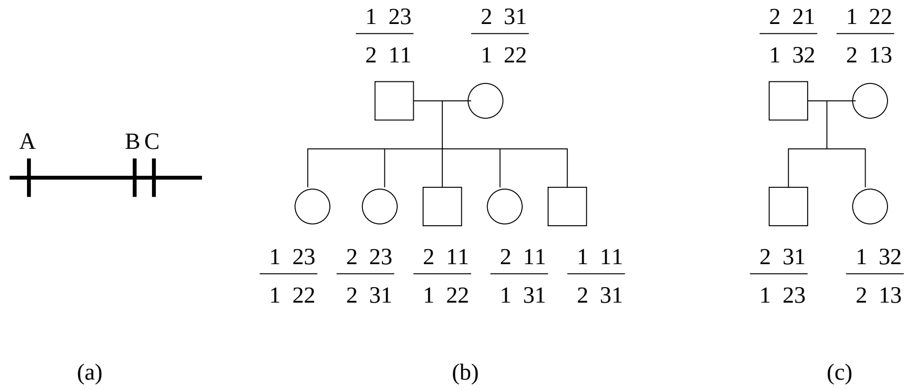

Figure 2.4: An example of genetic linkage between loci. (a) A map of three markers A, B, and C. Marker A is located distantly from the two adjoint markers B and C. (b)(c) Inheritance of the alleles in two pedigrees. The haplotypes are marked next to each individual. The haplotype above the horizontal line corresponds to the paternally inherited haplotype and the one below the line to the maternally inherited haplotype.

may be split into two or more segments in each meiotic step. In each step, the probability of two loci being inherited together is relative to their (genetic) distance.

The term *linkage disequilibrium* (LD) can be defined as nonrandom association of alleles at two or more loci. Consider two alleles A and B in different loci with population frequencies $p(A)$ and $p(B)$. When the alleles appear independently of each other in haplotypes, their expected frequency of occurring together in a haplotype is equal to $p(A)p(B)$. However, if the alleles appear together more often than expected, they are said to be in linkage disequilibrium, and its extent can be measured by pairwise linkage disequilibrium coefficient $\delta = p(AB) - p(A)p(B)$ [TeW98], where $p(AB)$ denotes the probability of the alleles appearing together in the same haplotype. In computation of LD, probabilities $p(A)$, $p(B)$, and $p(AB)$ can be estimated from the data. LD is generally computed for alleles at linked loci, but the definition also extends to loci at different chromosomes (nonsyntenic LD).

Linkage disequilibrium measures the extent at which alleles in the two or more loci are associated. On the other hand, the association of two alleles in the genotype at a single locus is measured by their departure from *Hardy-Weinberg equilibrium* (HWE). A population is said to be in HWE, if the genotype frequencies in a locus can be computed as products of the frequencies of the individual alleles. This condition is met if forces such as selection, mutation and migration have not changed allele frequencies over time [Wei96].

For example, consider two biallelic loci in different chromosomes. Let us assume that the frequencies of the two alleles A_1 and a_1 in the first locus are $P(A_1 = 0.1)$ and $P(a_1 = 0.9)$, and the frequencies of the alleles A_2 and a_2 in the second locus are $P(A_2 = 0.05)$ and $P(a_2 = 0.95)$.

If HWE is assumed, this yields population genotypic frequencies of $P(A_1A_1) = 0.01$, $P(A_1a_1) = 0.18$, $P(a_1a_1) = 0.81$, $P(A_2A_2) = 0.0025$, $P(A_2a_2) = 0.095$, and $P(a_2a_2) = 0.9025$. Since the genes are located in different chromosomes (i.e., the loci are unlinked), the population frequencies of the multilocus genotypes can be computed as products of the single-locus genotype frequencies. For instance, the frequency of the two-locus genotype $(\{A_1, a_1\}, \{A_2, a_2\})$ is equal to $P(A_1a_1) \cdot P(A_2a_2) = 0.0171$.

2.6 Genetic disease and genetic models

The genomes of all living organisms are constantly changed by *mutations*, which randomly alter the genetic information stored in the chromosomes of a cell. If a mutation occurs in a germline cell, it can be transmitted to the individual's offspring. If this is the case, the mutated allele becomes a part of the genome of the offspring, and he/she can in turn transmit it to the next generation. Under some conditions, an individual with even a single mutated base pair may develop a genetic disease. The target of disease gene localization is thus to find locations of the genes that contain the disease-predisposing mutations.

In disease gene localization, information about the disease statuses, or *phenotypes*, of the individuals are needed. Phenotype information is generally based on clinical or pathological observations, and it may give very little information about the underlying genetic factors of any individual. However, by studying the segregation of disease phenotype in pedigrees, some deduction can often be made about the underlying genetic architecture of the disease. By combining the phenotype and genotype data, localization of the disease genes may become possible.

Figure 2.5 illustrates the etiology of two types of genetic diseases. In a *monogenic disease* (Figure 2.5.a.) the disease phenotype is caused by a single gene, whereas the risk to develop a disease phenotype of a *complex* , or *multifactorial disease* (Figure 2.5.b.) can be controlled by multiple genetic, environmental and stochastic risk factors [Pri01].

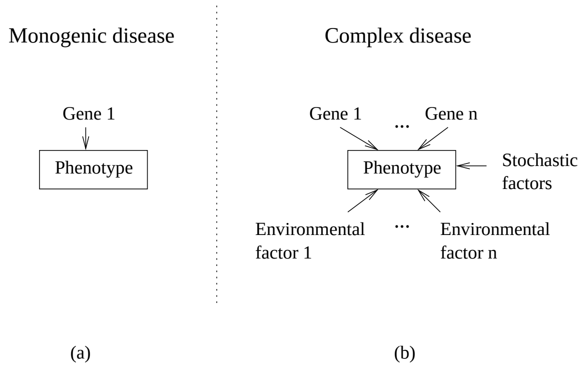

Figure 2.5: Factors contributing to the phenotype of (a) a monogenic disease and (b) a multifactorial disease.

Monogenic disease Recall that a human being has two copies of chromosomes $1, \dots, 22$. As a consequence, each (autosomal) gene is present in two copies, which gives rise to different *modes of inheritance* of genetic diseases. Monogenic diseases are the simplest family of genetic diseases. A *dominant* disease emerges if the individual has at least one copy of the mutated allele, while a *recessive* disease requires both alleles of a gene to be mutated. The *penetrance* of a genotype with respect to the disease gives a probability that a random carrier of the genotype is affected by the disease. If penetrance is equal to one, all carriers are affected.

Complex disease The phenotype of a *complex disease* depends on genetic, environmental, and stochastic factors. From a genetic point of view, a complex phenotype is often defined as a phenotype, where alleles in any single locus are neither necessary nor sufficient to cause the disease phenotype [Pri01].

The genetic component in a model of a complex disease can be complicated. First, multiple genes in different loci can contribute to the disease phenotype. In the simplest form of such *locus heterogeneity*, there are two disease loci in different regions of the genome, and a disease allele in either of these loci is sufficient to cause the disease.

Second, the disease phenotype can be caused by more than one allele in the disease locus. This phenomenon is referred to as *allele heterogeneity*. The different alleles may have a similar effect on the phenotype, or the magnitude of the effect may vary from allele to allele.

The locus and allele heterogeneity can exist simultaneously and form very complex phenotypes: several loci can contribute to the disease, and the combinations of different disease alleles in these loci can have different impacts on the disease phenotype.

In addition to the genetic component, the phenotype of a complex disease can depend on environmental factors. For instance, the emergence of a common complex disease, asthma, may be influenced by nongenetic risk factors such as cigarette smoking, presence of siblings, sex, and contact to farm animals [LNJ01]. To make things more complicated, responses to different environmental exposures may vary from individual to individual and be controlled by genetic factors, termed *xenogenetic susceptibility factors* [SSR95].

In this thesis, we refer to the individual loci that contribute to the disease phenotype as *disease loci*. In those loci, alleles that (either alone or together with alleles in other disease loci) increase the risk of an individual to be affected are referred to as *disease alleles*. The two forms of heterogeneity, locus and allele heterogeneity, are referred to as *genetic heterogeneity*. Due to locus heterogeneity, or environmental factors, individuals without any of the disease-predisposing alleles in a specific locus (or loci) can sometimes also become affected, in which case they are called *phenocopies* with respect to that disease locus (or loci).

Chapter 3

Motivation of the thesis

The purpose of this thesis is to create a simulation procedure for finding an optimal gene mapping strategy for a genome scan project in an isolated human population. The emphasis lies on methods that detect haplotype association. The work includes evaluation of the feasibility of different study strategies to mapping of human disease genes, and an analysis of how the factors of the population history (such as population size, age, and substructure) affect the power of a gene mapping study.

It is advantageous to conduct gene mapping projects utilizing the linkage disequilibrium present in *genetic isolates*, i.e., populations that have received only few immigrants from other populations in the past. Isolates have properties that can be of considerable advantage in the search of disease genes: depending on the size of an isolate, it may be reasonable to expect that only one — or at most a few — copies of a disease-predisposing allele has entered the initial population. Even if the population had expanded rapidly afterwards, a substantial amount of current-day diseased individuals would still be expected to carry a copy of the same mutation inherited from the original ancestor. The reduced degree of allele and locus heterogeneity, accompanied by the fact that the shared segments of ancestral haplotypes can be detected and they are longer among affected members of the isolates, can lead to easier detection of the predisposing genes compared to non-isolated populations [HDK92, Del93, DeW98].

Even though the genetic isolates offer advantages for study design, the extent and nature of the benefits depend on the population history of the isolate. For example, in a small population of constant size, the random genetic drift creates new linkage disequilibrium everywhere in the genome. As a result, there may be substantial differentation in marker allele frequencies at loci completely unrelated to the disease phenotype. Although this can be considered a nuisance, careful analysis of the process has lead to the development of drift mapping [TZL98] that utilizes this phenomenon.

The historical expansion rate is not, however, the only factor that has an effect on the genetic structure and heterogeneity of the current population. Factors such as initial population size, migration rate, population subdivision, and the number of founder mutations can well be expected to contribute to what we can see today. The effect of each of these factors can be approached analytically to some extent, but the complexity of the issue as well as the unknown effects of the events in population history may make theoretical computations difficult [ThN97]. There can be oversimplifications and errors in estimation of factors such as number of founder mutations, number of generations, disease gene frequency, marker mutation rates, and selection [Del93]. An alternative approach is to simulate the flow of genetic material and gather empirical evidence of the phenomenon under consideration. This approach has been to some extent implemented in studies such as [TZL98] and [DuF97]. In this thesis, we create a simulation framework that allows empirical analysis of the effects of population history as well as that of the intended study strategy. We are aware of the fact that the parametrization of the simulation process is still subject to errors and simplifications. However, we believe that these errors and simplifications are smaller than in the analytical approach, since the characteristics and events of the population history can be easier incorporated and expressed in more detail in the simulation approach.

In any genetic mapping study, it is desirable to estimate the power of the study strategy before the commencement of a sample collection. At this point, simulations can be of help. For genetic linkage analysis, where the intended data consists of (supposedly) independent extended pedigrees, the power of the setup can in principle be evaluated by repeatedly simulating the flow of alleles within each pedigree, computing a summary statistic (to detect linkage) for each pedigree, and combining the results. One must be cautious, since this requires assumptions about factors such as allele frequencies, genotyping success rates and disease models, but the results can still be applicable in practice. For association analysis, where the samples are collected at population level, one must be even more cautious: the analyst must make assumptions about all the aforementioned parameters and, in addition, the characteristics of the population history, which are expected to have a crucial impact on what we can eventually see in the data. It has been stressed that the variation of events that are functions of a population history is generally very large [ThN97], which may make it very difficult to obtain accurate power estimates.

Now, an important question arises: which demographic factors are required to accurately simulate the history of the population and the transmission of genetic material within it. In Finland, for example, detailed demographic records and statistics are available from the early 18th century onwards, whereas the earlier estimates of the parameters must be obtained from less covering sources, such as Swedish tax records [KMN94]. Even if one concentrates on one of the recently founded isolates of our country (such as Kainuu), this creates uncertainty to the initial setup of the study population. One might anticipate that the inaccuracy of some parameters are expected to have a larger impact on the current population than that of others. One of the objectives of this thesis is to recognize these parameters, and analyze the mechanism in which each of the parameters may affect or alter the genetic data.

Among the considerations about the population history, a part of the design of study strategy in any genetic mapping project are the decisions of how the project is divided into subphases, what kind of marker maps are used and which kind of results are considered prominent in each case. As an aid to these considerations, the target is to present a method for finding the minimum cost for a multiple-phase mapping study.

Still another important aspect in disease gene localization by haplotype association is how to measure the linkage disequilibrium between the disease and marker loci. For this purpose, a collection of measures of linkage disequilibrium have been proposed, and a theoretical analysis of their properties can be found in [Guo97]. The most widely used measures do not extract all the information from the allele frequencies. Instead, they consider each locus biallelic, lumping together all alleles except one. Lewontin considers the concept of linkage disequilibrium ill-defined and has demonstrated [Lew88] that there is no generally gene frequency-independent linkage disequilibrium measure, but there may be measures that perform well under particular conditions. The applicability of these measures, and some of their multiallelic counterparts, are analyzed in the context of haplotype analysis.

To summarize, the main contributions of this thesis are:

1. To create a software package for simulating population history and inheritance of chromosomes within the populations. This includes development of an algorithm for simulating crossovers and chiasma interference.

2. To create a computational procedure and a software implementation for finding the optimal gene mapping strategy for a genome scan in an isolated population, subject to assumptions about the underlying true disease model.

3. To analyze the effect of different characteristics of population history on the feasibility of a genome scan.

4. To compare the applicability of commonly used association mapping methods in the context of a genome scan carried out in a population isolate.

The population simulator developed as a part of this thesis has been used in the development phase of the methods published in [TOV00], [TOV00b] [STO01], and [OOS02], and applied in [HVS99] and [NMH01]. Likewise, participation in preparing these works has had a major impact on the development of the simulator software.

Chapter 4

A population simulator

4.1 Introduction

The interest in disease gene mapping has shifted towards complex diseases, where the relationship between the phenotype and the genetic and environmental factors can be complicated. As a consequence, the analytical computations may become too complex, which in turn makes simulation methods an attractive choice [Gau95, Bal01]. Simulation methods have been previously developed for both pedigree and population data.

For pedigree data, SLINK simulation package [Ott89, WOL90] is probably the best known simulator software, which is still frequently used for power calculations of parametric linkage analysis (see Section 5.1.1). SLINK creates a probability distribution for the genotypes of each individual conditioning on the given phenotypes and the genotypes already allocated to other individuals. Whereas SLINK concentrates on simulating genotypes at marker loci, Genetic Analysis Package (GAP) software [Gau95] uses Gibbs sampling to simulate genotypes at the disease locus as well as values of environmental covariates and a polygenic component.

POPSIM software [HWS98] is one of the software packages designed for creating population genotype data. It allows incorporating some population parameters (population size, subpopulation structure and number of generations), multiple unlinked disease loci and a complex disease model. The software is suitable for analysis of the influence of factors such as genetic drift, recombination and admixture, and can be used as a tool in studying the patterns of linkage disequilibrium and data collection issues. The software can handle large populations, but the inheritance data obtained is restricted to the contents of the predefined disease and marker loci [HWS98].

EASYPOP simulator software [Bal01] is designed for creating genotypes for a wide variety of different organisms, and it offers a large choice of breeding and population models. The software, which is oriented towards population genetics purposes, allows simulation of multilocus genotypes, but the incorporation of complex disease models and family structure required in the analysis of human genetic are not supported.

In this work we have designed and implemented methods and a software package for the simulation of human genetic data in isolated populations. The software allows simulation of the entire process of population formation, inheritance of chromosomes, phenotype definition based on complex disease models, collection of individuals, and generation of marker data. To overcome some of the limitations of [Gau95] described above, the full history of the chromosomal segments inherited IBD is stored for all individuals in the population. As a consequence, repeating the analysis with different marker maps is trivial and does not require heavy re-simulation of the generation of the population and inheritance of chromosomes.

The software package developed simulates the flow of genetic material within pedigrees created in accordance with user-specified population histories. By repeating the inheritance within a fixed population history, it is possible to obtain a large collection of artificial data sets. These data sets, in turn, can be used to assess the significance of an observed finding, to estimate the power of the proposed study setup, to analyze the connection between disease models and population history, and to perform comparative studies on different methods and measures used in disease gene localization. These tasks will be discussed in the forthcoming sections.

The components of the simulation package are shown in Figure 4.1. There are four components altogether: a pedigree generator, a chromosome simulator, a sampler and a marker tool. First, the description of the population parameters is given to the pedigree generator. It creates an artificial pedigree that corresponds to these parameters. This pedigree is again given to the chromosome simulator, which simulates the reproduction events in the pedigree and creates the list of each individual's chromosomal segments as output. Next, the desired disease model and sample sizes are given, and the computer extracts the segment data of corresponding individuals. At this phase, the segment data together with the desired marker map specification is passed to the marker tool, which finally produces artificial marker data.

The artificial marker data is used to perform the analyses (typically gene localization), after which the entire process is repeated the desired number of times to obtain an empirical distribution of the outcome of the analyses. In the next subsections, each of the four components involved will be examined in more detail. For different purposes, alternative versions of

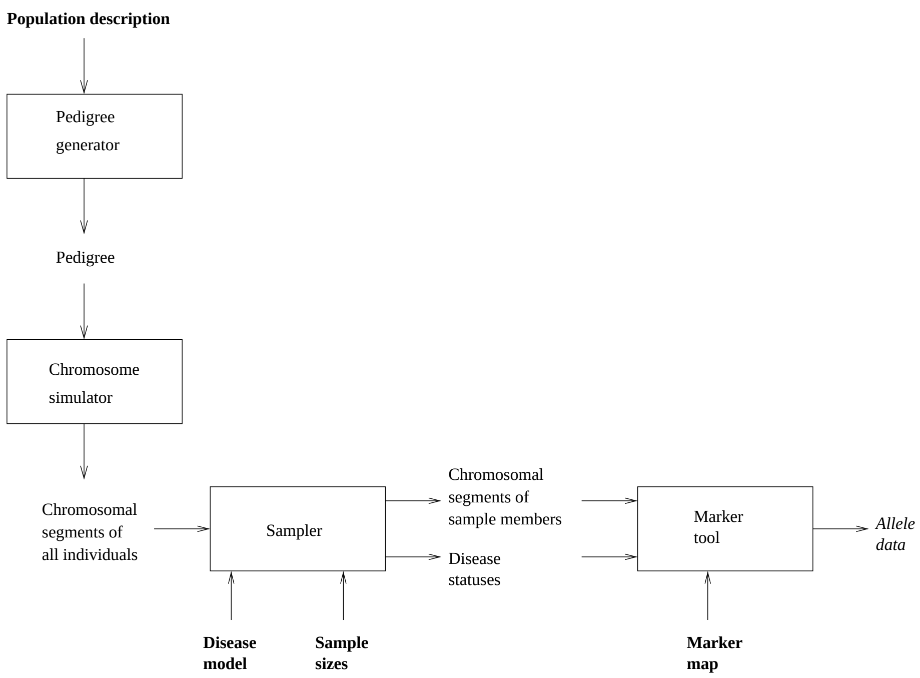

Figure 4.1: Components and data flow of the simulation software in a typical context of creating case/control data for analysis of a complex disease. The four components of the simulation software are displayed in the boxes. The data includes the input data given by the user (bold font), the output data of the simulation system (italic font) and the intermediate data passed to another component of the software (regular font). The arrows indicate the flow of information between the user and the software as well as between the software components.

some modules have been implemented.

There are many computational problems involved in the simulation of large populations. We have to be able to generate a large pedigree, and to mimic the recombinations of chromosomes, etc. Our problem area is efficient implementation of the methods, but even more important is the correctness of the specification. That is, we have to be convinced that the methods used in simulation correspond as closely as possible to the real phenomena. For example, producing a sample population that fits together with the observed population history is by no means simple.

4.2 Simulating population history

The first component of the simulation software, the population generator, is used to produce artificial pedigrees that are of the desired shape and structure. The main use of the generator is to create pedigrees that cover the history of a recently expanded population isolate for the past few centuries. In the beginning, an initial population is introduced, after which the members of the population are allowed to mate, give births and, eventually, die. The population may also be reinforced by incoming immigrants. The final outcome of such a process is the expanded contemporary population. If the historical parameters have been specified accurately, the real-life population of the isolate can be seen as one possible outcome of that process.

The basic structure of the population simulator is simple. We form the pedigree by starting from founders, simulating births, deaths, and immigration, all according to the specified parameters. From the viewpoint of computational complexity, we aim at a linear time implementation of the population generator. This means that we have to avoid doing complex searches within the generation of the pedigree.

There are two versions of the population generator: one handling overlapping generations and complex population history events, and another which divides the time scale into distinct generations and assumes steady exponential growth, but allows population substructuring within each generation. The first implementation can be used to accurately describe the desired population history, including factors such as annual changes in population size, death age distributions, and immigration rate. The other implementation only includes a small number of parameters, but allows analyzing the effect of this limited set of parameters (population size and age, substructure, and immigration rate) efficiently without the needing to incorporate complex data about the other factors of less interest.

4.2.1 Homogeneous populations, overlapping generations

The first version of the population generator generates pedigrees based on the assumptions of random mating and absence of admixture within the population, but allows a wide variety of demographic parameters. These parameters are listed in Table 4.1. Examples of population descriptions that can be utilized by the simulator are summarized in Figure 7.1 (on page 95). For each parameter, the number of intervals and their endpoints (including year ranges and age groups) can be freely specified by the user. Moreover, this can be done independently for both sexes. The parameters are selected to facilitate simulation of relatively young population isolates with possi-

foundation year	The year when the initial population is created.
ending year	The year when the simulation ceases.
population size*	Number of individuals alive.
age structure*	When applied to the foundation year, defines the age structure of founder individuals. When defined for any later year, specifies the age structure of the incoming immigrants.
death risk*	Probability that a living person will die during a period of one year. The death risk is defined for all ages and should eventually reach unity.
age of marriage*	The age at which a living individual is taken into account when organizing marriages.
maximum age of pregnancy*	The highest age at which a female individual may give birth.
immigration rate*	The rate at which immigrants arrive. The rate is proportional to the overall size of the living population.

Table 4.1: Parameter explanations of the population generator for overlapping generations. The name of each parameter is given on the left and a short explanation on the right. An asterisk (*) indicates that the parameter can be defined as a function of time.

ble fluctuations in immigration and mortality during history. Note that emigrating individuals do not contribute to the gene pool of the population, which makes it possible to incorporate the effect of emigration into the death risk parameter. Thus, the emigration rate is not included as an individual parameter in the implementation.

Algorithm 1 gives an overview of how the population generator is implemented. The creation of the initial population is described in Algorithm 2, the management of the deceased individuals in Algorithm 3, introduction of marriages in Algorithm 4, and, finally, creation of new individuals (newborn or immigrants) in Algorithm 5.

Generation of the initial population The initial population consists of the desired number of unrelated individuals, all labeled as founders. The age structure of the individuals is defined in the population description, so every individual is assigned a birth year, dating back to the time before the starting year (with the exception of infants, whose birth year equals the starting year). It will also be decided whether an individual will get married during his or her lifetime and, if so, the year when that will happen. The sexes of founders are not yet relevant and remain unspecified at this time. Once the population has been created, its members are allowed to form couples. This is done by pooling all individuals that are considered fit for marriage and have just reached, or passed beyond, their marriage year. Within this pool as many couples are formed randomly as possible, and the sex of each member of a couple is defined in such a way that every couple will consist of one male and one female. After a couple is formed, its members will in the future be considered to be possible parents for children.

Annual operations When the initial population is formed, the simulation cycle is repeated every year. First, the targeted population size for the current year is computed according to the population description. After this, it is decided who can carry on living and who will face his/her death that year. Individuals assigned dead are removed from the current population.

When the dead persons are removed, the next phase is to find those individuals who have reached their marriage age. The procedure is similar to that described above for the initial population: the individuals are pooled, couples formed and sexes defined.

Next, new individuals are created. These individuals consist of both newborns and immigrants. Let us denote the targeted population size for year i by $s(i)$, the number of deaths occurred that year by $d(i)$, and the immigration rate by $m(i)$. The net population growth for an arbitrary year y is then computed from expression $s(y) - s(y-1) + d(y)$, and this is allocated to newborns and immigrants so that the number of immigrants is $m(y)(s(y) - s(y-1) + d(y))$ and the number of newborns $(1 - m(y))(s(y) - s(y-1) + d(y))$. When these numbers are computed, the corresponding individuals are created in a similar way as with the initial population: the new individuals get a birth year, which in case of a newborn is the current year, and, in case of an immigrant is taken from the age structure specified in the population description file. The individuals also obtain a possible year of marriage but the sex remains unspecified. Immigrants are labeled as founders and are not linked to any existing members of the pedigree. However, all newborns are considered to be children of existing couples. For each newborn, a random couple is selected from the set of such couples where both members are alive and the mother has not reached her maximum age of having children yet.

At first glance, it can seem an odd way to select the parents after the children have been created instead of creating a predefined number of chil-

Algorithm 1 Creating pedigrees in overlapping generations.
Constants: see Table 4.1. Notation $parameter[y]$ denotes the parameter value in year y

$indmin := 0$
$indmax := 0$
$create\_initial\_population()$ {See algorithm 2}
for $y = foundation\_year$ to $ending\_year$ **do**
  $s[y] := population\_size[y]$
  $d[y] := death()$ {See algorithm 3}
  $g[y] := immigration\_rate[y] * (s[y] - s[y-1] + d[y])$
  $b[y] := (1 - immigration\_rate[y]) * (s[y] - s[y-1] + d[y])$
  $arrange\_marriages()$ {See algorithm 4}
  $freeslot := 0$
  for $i := 1$ to $num\_couples$ **do**
    if $ind[couple[i].male].death\_year = 0$
    and $ind[couple[i].female].death\_year = 0$
    and $y - ind[couple[i].female].birth\_year \leq$
    $maximum\_age\_of\_pregnancy$ **then**
      $couple[{+}{+}freeslot] := couple[i]$
    end if
  end for
  $num\_couples := freeslot$
  for $i := 1$ to $b[y]$ **do**
    $ind[{+}{+}indmax] := create\_individual(y, NEWBORN)$ {See algorithm 5}
  end for
  for $i := 1$ to $g[y]$ **do**
    $ind[{+}{+}indmax] := create\_individual(y-$ random$(age_structure[y]), IMMIGRANT)$ {See algorithm 5}
  end for
end for
Output all individuals in memory

Algorithm 2 *create_initial_population*()

```
for i := 1 to MAXAGE do
  n := population_size[foundation_year]*
  age_structure[foundation_year][i]
  for j := 1 to N do
    ind[++indmax] := create_individual(foundation_year − i,
    NEWBORN) {See algorithm 5}
  end for
end for
```

Algorithm 3 *death*()

```
num_dead := 0
for i := indmin to indmax do
  if death_risk[year] > random [0, 1] then
    ind[i].death_year := year
    Output individual i
    num_dead++
  end if
end for
while ind[indmin].death_year > 0 do
  indmin++
end while
return num_dead
```

dren for each couple. However, the latter approach would have been difficult to do correctly, since at the time of couple formation, it may be difficult to find out what the correct number of children for each couple would be as the birth rates and death risks change as functions of time, and it may be possible that one of the intended parents dies prematurely. The selected approach leads to a much simpler implementation, always results in a correct number of offspring, and, in the special case of constant population size, will result in a similar outcome, as will be shown.

Consider a couple that is fertile for a period of y years. For simplicity, let us assume that the population stays constant in size for this time, and c children are born each year. Furthermore, let us denote the number of fertile couples at any timepoint by n. If the population size remains constant, numbers c and n are also expected to stay more or less constant. The childbirths can be considered discrete events in a continuous time scale, and thus be modeled as a Poisson process, where the number of children

Algorithm 4 *arrange_marriages*()

```
numcand := 0
for i := indmin to indmax do
  if ind[i].marriage_age ≤ year − ind[i].birth_year and ind[i].sex = 0
  then
    candidate[++numcand] := i
  end if
end for
for i := 1 to numcand do
  swap :=random {1, ..., numcand}
  temp := candidate[i]
  candidate[i] := candidate[swap]
  candidate[swap] := temp
end for
for i := 1 to ⌊numcand/2⌋ do
  couple[++num_couples].male := candidate[2i − 1]
  couple[num_couples].female := candidate[2i]
  ind[couple[num_couples].male].sex := 1
  ind[couple[num_couples].female].sex := 2
end for
```

Algorithm 5 *create_individual*(*birth_year*, *type*)

```
if type = IMMIGRANT then
  ind.birth_year := random(age_structure[y])
  ind.father := 0
  ind.mother := 0
else
  ind.birth_year := year_of_birth
  c := random {1, ..., num_couples}
  ind.father := couple[c].male
  ind.mother := couple[c].female
end if
ind.sex := 0
ind.marriage_age := random(age_of_marriage[ind.birth_year])
return ind
```

allocated to the specified couple follows the distribution $P(yc/n)$. This selection of parameter provides the desired expected value for the number of newborns.

On the other hand, consider the case where a parental couple is randomly selected for each newborn, as is done in the software described here. The probability that a given couple obtains a specific child is equal to $1/n$, and there are altogether yc such possibilities. The distribution for the number of children X that a couple will obtain now has distribution $X \sim Bin(yc, 1/n)$, which is accurately approximated by distribution $P(yc/n)$, when yc is large and $1/n$ small. This is generally the case, so the selected procedure has approximately a similar outcome as allocating a number of children to each couple based on Poisson distribution.

Since the death risks in different age groups are specified by user, the way to reach the desired population size is to control the number of births. If the user specifies a very rapid decrease in populations size without increasing mortality as much as necessary, it may be possible that the population size does not sink fast enough. This is the only case that the observed population size does not match the one given by the user, and is nearly always a result of misspecification in the population history.

Software implementation The software has been implemented in the C programming language. In processing the population description file we utilize the Lex lexical analyzer [Les75] and Yacc compiler-compiler [Joh75]. Use of these tools allows easy format for description of parameters as function of time, age group and sex.

The time requirement of the chromosome generator is approximately linear with respect to the number of individuals created. This linear time requirement is verified by the results of Figure 4.2. On the left, the execution times on an AMD Athlon 1,333 MHz processor are plotted as a function of the size of a final population. In the simulations, we used exponentially growing populations of constant age and population BASIC (see Figure 7.1) with the modification that the population size was set to grow exponentially from the initial value of 300 to its final value in a period of 500 years. On the right, the execution times are shown as a function of ending year, this time fixing the size of the final population to 100,000. If the final population size is fixed, the execution time behaves linearly as a function of the number of years. However, in an exponentially growing population, the execution time thus is expected to grow exponentially as a function of years, which is also a lower bound dictated by the need to output every individual in the pedigree. Fortunately, this is not a problem for simulation of young population isolates, where the final population size is expected to be in the ballpark of a few hundred thousand individuals at most. For example, for the Finnish population isolate of Kainuu, the final population size is less than 100,000 [Sta02], and the execution time well below one minute.

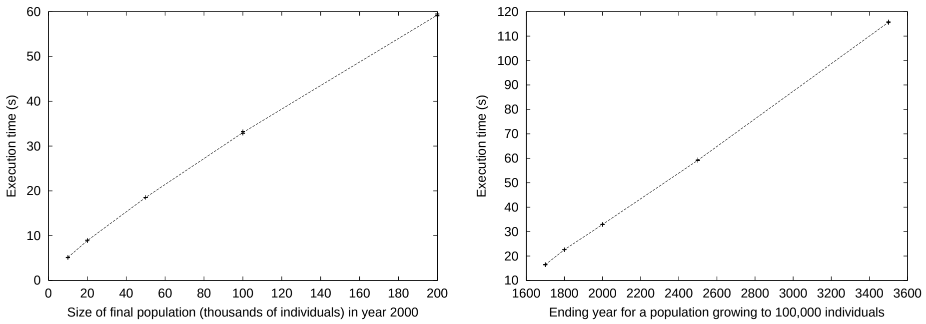

Figure 4.2: Execution times on an AMD Athlon 1,333 MHz computer for population BASIC, assuming annual exponential population growth starting from 300 individuals. (A) A constant number of years (500) and a variable size of the final population. (B) A variable ending year and a fixed size of final population (100,000). In the figures, the pedigree generation was repeated five times for each parameter value. The line denotes the average execution time.

Since the maximum age of an individual and the minimum age for giving birth are limited, there is in practice a maximum number of ancestors of any given individual that can be alive simultaneously. Consequently, the number of individuals required to keep in the memory at any given time does not exceed the number of living individuals in the population multiplied by a small constant. As the parents cannot be discarded before the birth of their children, the space requirement of the population generator is proportional to the current population size. In an exponentially growing population, the space requirement grows exponentially as a function of the time. In the pseudocode representation (Algorithm 1) and in the implementation, all individuals are stored in a global array. If necessary, the use of variables *indmin* and *indmax* would easily allow using the table of individuals in a cyclic way, saving some space.

4.2.2 Substructured populations, distinct generations

Hidden population substructure may cause large fluctuations in the frequencies of genetic variants [ThN97]. To analyze the effect of such substructure, we desire pedigrees with spatial – or more likely socio-spatial –

substructure. For simplicity, and to avoid making the previously described population simulator too complex, we restrict the analysis of substructure to exponentially growing populations. This less complicated version of the population generator was created in *awk* programming language to produce exponentially growing, substructured pedigrees in distinct generations.

There is internal migration between the subpopulations, and also immigration from outside the system can be allowed. Individuals born to a high-quality subpopulation emigrate to adjacent subpopulations with low probability. In each generation, the parents are divided into random couples, and a random parental couple is selected for each individual. If there is only one subpopulation with M parental couples and N offspring, then the number of children of a specific couple follows binomial distribution with parameters M and $1/N$, and their expected number is N/M. In case of more than one subpopulation, the parent selection is done separately in each subpopulation.

The study population is represented as a matrix, where each cell has a value p_{ij} denoting its *quality*. The quality of a subpopulation means the inverse of the probability that an individual born in one of the populations migrates to an adjacent population before fertility. Thus, internal migration is controlled by the cell quality values p_{ij}: each individual born in cell (i, j) migrates to any of the adjacent subpopulations before having offspring with probability $1/p_{ij}$. Subpopulations and migration probabilities are illustrated in Figure 4.3.

The function of this version of population generator is described in Algorithm 6. Algorithm 7 describes creation of new individuals, Algorithm 8 selection of parents to the newborn, and Algorithm 9 migration between the subpopulations.

During the execution, it is only necessary to store the last two generations in the memory at each time. Both the time and space requirements are linear with respect to the number of individuals in the current generation, which means that for an exponentially growing population, these requirements grow exponentially, albeit rather slowly in case of a small isolate.

4.3 Simulating inheritance of chromosomes

The purpose of the chromosome simulator is to simulate the propagation of genetic material in a given pedigree, which can either represent a (very large) population or a single family of a few individuals. The simulation of inheritance is done by simulating crossovers in each meiosis, and maintaining a list of each individual's chromosomal segments related to the original

Algorithm 6 Generating substructured exponentially growing populations in distinct generations.

Constants: G: number of generations
S_0, S_G: sizes of the initial population and the population in generation G
B: number of subpopulations
$p[i]$: quality of subpopulation i

$S_0 := 0$
$f := (S_G/S_0)^{1/G}$ {Growth rate}
for $g = 1$ to $G + 1$ **do**
  $S_g := S_{g-1} f$
  $immigrants = I(S_g - S_{g-1})$
  $newborn = S_g - immigrants$
  for $i = 1$ to $newborn$ **do**
    $ind[i] = create\_individual(i)$ {See algorithm 7}
  end for
  for $i = 1$ to $newborn$ **do**
    $select\_parents(i)$ {See algorithm 8}
  end for
  for $i = 1$ to $immigrants$ **do**
    $ind[newborn + i] = create\_individual(newborn + i)$ {See algorithm 7}
  end for
  Output parental generation $g - 1$
  $migrate()$ {See algorithm 9}
  for $i = 1$ to S_g **do**
    $prev\_ind[i] := ind[i]$;
    for all $i = 1$ to B **do**
      for all $j = 1$ to S_g^i **do**
        $prev\_index[i][j] = index[i][j]$
      end for
    end for
  end for
end for

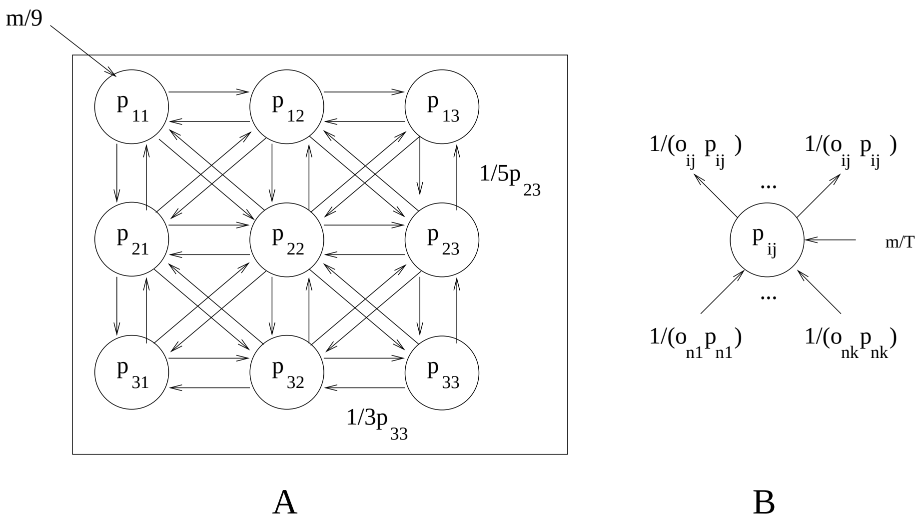

Figure 4.3: Migration in the population. (A) A population consisting of 3×3 subpopulation structure. Arrows between subpopulations represent internal immigration. The probability of an individual to migrate to a neighbouring subpopulation is the inverse of the cell quality value p_{ij}. The immigrants are distributed uniformly between the nine subpopulations. (B) Immigration and emigration for a single subpopulation (i, j). The subpopulation quality is denoted by p_{ij} and the number of subpopulations adjacent to subpopulation (i, j) by o_{ij}. The total number of immigrants is indicated by m and the number of subpopulations by T. The adjacent subpopulations are indicated by $n1$ to nk.

Algorithm 7 create_individual(i)

create ind[i]
$ind[i].sex$:="unknown"
$ind[i].father := 0$
$ind[i].mother := 0$
$b :=$ random$\{1, 2, \ldots, B\}$
$ind[i].subpopulation := b$
$S_g^b + +$
$index[b][S_g^b] := i$

Algorithm 8 select_parents(i)

$b := ind[i].subpopulation$
repeat
  $p_1 := prev\_index[b][\text{random}(1, S^b_{g-1})]$ {Select father}
  $p_2 := prev\_index[b][\text{random}(1, S^b_{g-1})]$ {Select mother}
until $prev\_ind[p_1].sex \neq$ "female" and $prev\_ind[p_2].sex \neq$ "male" and $p1 \neq p2$
if $prev\_ind[p_1].sex =$ "unknown" and $prev\_ind[p_2].sex =$ "unknown" **then**
  $prev\_ind[p_1].sex :=$ "male"
  $prev\_ind[p_2].sex :=$ "female"
  $prev\_ind[p_1].spouse := p_2$
  $prev\_ind[p_2].spouse := p_1$
  $ind[i].father = p_1$
  $ind[i].mother = p_2$
else if $prev\_ind[p_1].sex =$ "male" **then**
  $ind[i].father = prev\_ind[p_1]$
  $ind[i].mother = prev\_ind[p_1].spouse$
else
  $ind[i].father = prev\_ind[p_2].spouse$
  $ind[i].mother = prev\_ind[p_2]$
end if

Algorithm 9 migrate()

for $i = 1$ to S_g **do**
  $r = 1/p[ind[i].subpopulation]$
  if $r <$ random[0,1] **then**
    $ind[i].subpopulation :=$ random adjacent subpopulation
  end if
end for

ancestors. During the simulation, the ancestral origin of each segment of every individual is kept track of. Such an approach has been proposed for small, known pedigrees by Terwilliger et al. in [TSO93]. Here, the method is applied on a much larger scale for population isolates ranging to hundreds of thousands of individuals.

Besides human beings, the chromosome simulator can potentially be used to simulate the inheritance in any diploid organism, including plants and animals.

Input and output data The simulator is given an input file describing the structure of the pedigree. This pedigree file can be obtained from the pedigree generator, or, alternatively, the user may prefer entering a real, known pedigree structure. This description of the pedigree includes each individual's identification number and sex as well as references to his/her father and mother.

The user also enters the desired genetic length of the chromosomal region to be simulated. Different genetic lengths can be given to different sexes. (In spite of some evidence for longer genetic maps for females [WSL89], the sex-specificity of chiasma interference could not be verified in human beings in a recent study of Broman and Weber [BrW00]). In addition, the user can separately mark those individuals whose data should be included in the output of the simulator.

Figure 4.4 demonstrates the information processed by the chromosome simulator. The structure of the pedigree is described in the input file, and as an outcome of the simulation process, the simulator provided a list of chromosomal segments that the desired individuals (in the example individuals numbered 8 and 9) have inherited from the founder individuals. The start and end points of each segment as well as the segment's ancestral origin is provided.

Implementation details The function of the simulator is described in Algorithm 10. The implementation of meiosis is depicted in further detail in Algorithm 11. Whenever a new individual is read, the simulator checks, whether he or she is a founder. If so, two (or one, if only one of the parents is a founder) new, continuous chromosomal segments $[0, 1]$ are created. In the example of Figure 4.4, this is done for individuals 1, 2, 3, and 4. In the opposite case, the segment data of the parents is retrieved, meiosis is simulated in both parents, and the resulting chromosomal segments, possibly split by newly introduced crossover events, are copied to the offspring. Internally, the chromosomal data of each individual is represented as a linked list of chromosomal segments referring to the original founder who has brought that segment into the population. The paternally inherited chromosome of individual 5 has undergone one recombination event while her maternally inherited chromosome has been passed intact. As generations pass, the segments of each individual can always be traced back to their original founder sources, even though the genetic length of any single segment gradually diminishes. From individual 8, for example, we observe that her paternally inherited chromosome consists of three continuous founder segments: the first 28.6 cM are inherited from the maternally inherited chromosome of founder 1, the next 46.3 cM from the paternally inherited chromosome of the same founder, and the last 25.1 cM from the paternally inherited chromosome of founder 2. Similar data is available about the other chromosome of individual 8 as well as both chromosomes of the other requested individual, number 9.

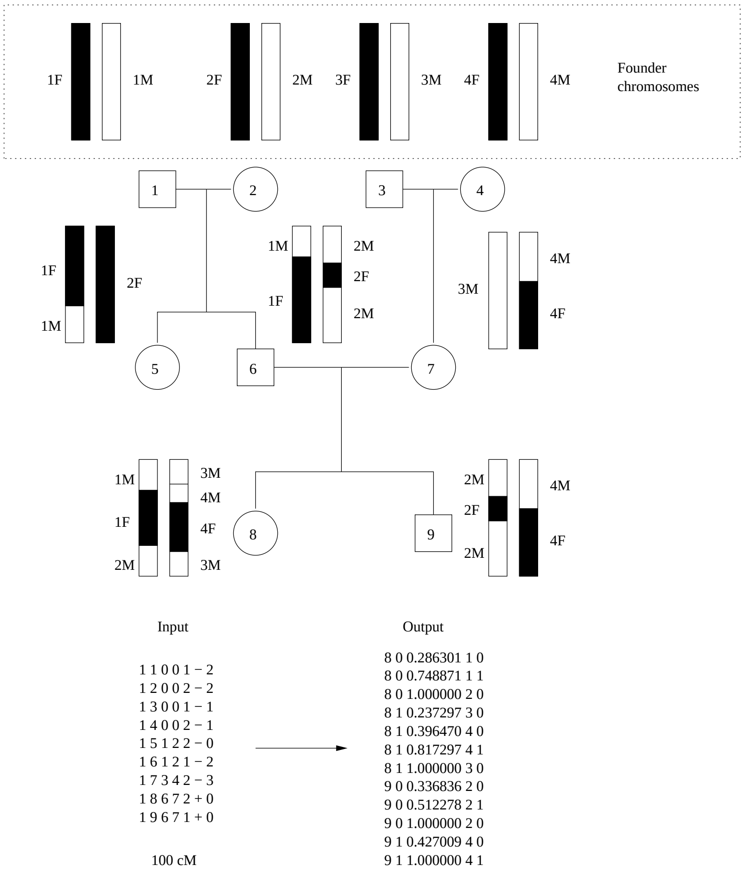

Figure 4.4: An example of inheritance of chromosomal segments in a pedigree of 9 individuals. For a full description, see page A–1 of Appendix A.

Meiotic models The meiosis is implemented using two alternative models, one with chiasma interference and one without. In the latter model, previously described in Algorithm 11, the number of chiasmata in the chromatid passed to the offspring is Poisson distributed with intensity equal

Algorithm 10 The chromosome simulator

repeat
  Read data of individual i: $ind[i].father, ind[i].mother, ind[i].sex$
  if $ind[i].father = 0$ **then**
    $C_f[i] := \{[0, 1, i, f]\}$ {New founder chromosome}
  else
    $C_f[i] := meiosis(ind[i].father)$ {See algorithm 11}
  end if
  if $ind[i].mother = 0$ **then**
    $C_m[i] := \{[0, 1, i, m]\}$ {New founder chromosome}
  else
    $C_m[i] := meiosis(ind[i].mother)$ {See algorithm 11}
  end if
until End of file

to the length of the chromosomal region in Morgans. This corresponds to the definition of Morgan itself as a unit of genetic distance. Having determined the number of crossover events in the chromatid, their locations within the region are taken from uniform distribution. This corresponds to the use of Haldane's mapping function, where the recombination fraction θ is converted to genetic length x (in Morgans) by use of formula

$$x = -\frac{1}{2}\ln(1 - 2\theta).$$

Even though it is disputable whether chiasma interference exists in human beings, an alternative model that takes it into account has been implemented in the software. However, due to this uncertainty, chiasma interference is not assumed in the experiments of subsequent chapters. In the model implemented, the possible first crossover event takes place in a random location in the chromosomal region. When subsequent chiasmata are formed, the location within the chromosomal region is no more uniformly distributed. Instead, the probability of a chiasma being formed in location x_i is the smaller the closer x_i is to any one of the previous chiasmata.

Implementation of chiasma formation We developed a method to simulate meiosis with desired degree of chiasma interference. In the method, the strength of chiasma interference is controlled by a user-specified parameter β. To define the location x_i of next chiasma in a chromosome of length l, pairs (x, y), where x and y are random real numbers in interval $[0, 1]$, are repeatedly created. Each time, it is checked whether the family of inequalities

Algorithm 11 meiosis($parent$)
Constants: $genetic\_length$: size of the region in Morgans
Subroutines: $segmstart(i, j)$, $segmend(i, j)$, $founder(i, j)$, $fdr\_phase(i, j)$ return the starting point, ending point, founder identifier and paternal phase of the jth segment of $parent$, arranged by location in the region.
$Poisson(i)$ returns an integer randomly sampled from Poisson distribution with parameter i

```
C' := {}
count[f] := 0
count[m] := 0
crossover_points := Poisson(genetic_length)
curr_phase := random {m, f}
for i = 1 to crossover_points do
  cp[i] := random [0, 1]
end for
cp[crossover_points+1] := 1
curr_cp = 1
startpoint = 0
repeat
  repeat
    repeat
      count[curr_phase] + +
    until segmstart(curr_phase, count[curr_phase]) ≤ startpoint
          and segmend(curr_phase, count[curr_phase]) > startpoint
    endpoint = max{segmend(curr_phase, count[curr_phase]), cp[curr_cp]}
    C' := C'∪[startpoint, endpoint, founder(curr_phase, count[curr_phase]),
          fdr_phase(curr_phase, count[curr_phase])]
    startpoint := endpoint
  until endpoint = cp[curr_cp]
  if curr_phase = m then
    curr_phase := f
  else
    curr_phase := m
  end if
  curr_cp + +
until curr_cp = crossover_points + 2 {End of chromosome}
Return C'
```

$$\forall j \in \{1, \ldots, i-1\} : y \leq 1 - e^{-\beta l |x - x_j|}$$

holds, and, if so, the value of x above is taken to be the new location of the chiasma, x_i. Otherwise, a new (x, y) pair is created.

To illustrate the procedure of determining locations for chiasmata, consider the example of Figure 4.5, where chiasma formation is illustrated for a chromosome of 100 cM. In this example, two previous chiasmata have been formed at genetic locations 40 and 62 cM from the telomere. With a value of interference parameter $\beta = 20$, two probability curves departing from existing crossover points (see x-axis) are drawn. The overall acceptance probability curve is the minimum of these two curves, depicted with a dashed line. When a subsequent chiasma is going to be formed, a random point in the $[0, 1] \times [0, 1]$ plane is selected, and, in case the point is located below the dashed acceptance probability curve, its x-coordinate is taken to be the next location for a chiasma. If the point happens to be above the curve, a new point is selected, and this procedure is continued until the desired number of locations for chiasmata have been formed. This approach guarantees that the number of chiasmata is similarly distributed as in the no-interference model, the only differences being in their locations. It should be noted that when $\beta \to \infty$, the procedure corresponds increasingly accurately to the no-interference procedure.

Time requirements The data structure for each individual in memory contains a linked list of the segments of founder chromosomes. In each meiosis, the list is expected to get slightly longer: the very definition of genetic length dictates that for a region of m Morgans (m constant), the expected number of crossover events in the meiosis is equal to m. The chromosome, when passed to the offspring, contains all the original crossover points and the m new crossover points introduced in the last generation. As a consequence, the expected number of ancestral segments of an individual grows by m, which is constant. The expected number of ancestral segments of an individual living g generations after the foundation is thus proportional to g. The list of segments must be kept in memory for each potential parent, since the list needs to be scanned in the meiosis of that individual. For a population of n individuals, the number of potential parents is always less than cm, where c is a constant that follows from the fact that the number of living ancestors of an individual is always limited, the exact value of c depending on the death risk distribution.

As a consequence, the total number of segments in memory is a linear function of the number of individuals n in the generation multiplied by

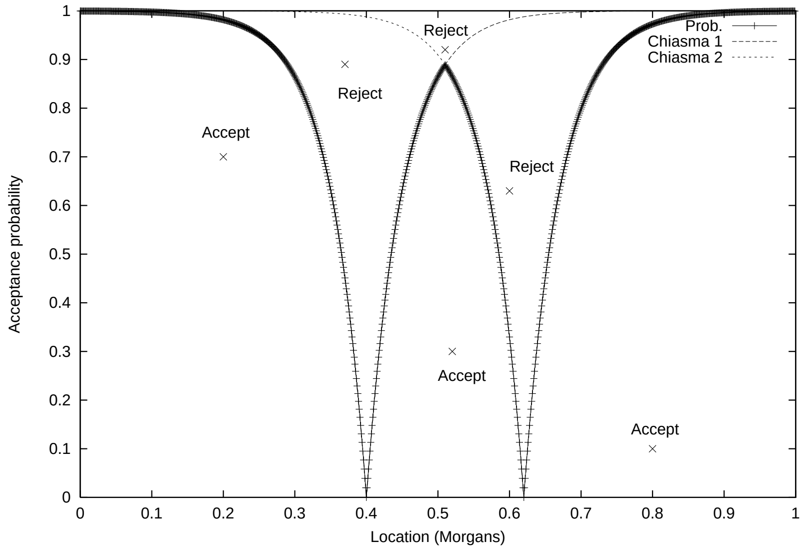

Figure 4.5: Determining the location of the next crossover. In this example, two previous crossovers have been formed at points 0.4 M and 0.62 M. The probability of next crossover at these points is zero, and increases when departing from the immediate surroundings of these points. The rate of increase is controlled by the interference factor β, which has the value of $\beta = 20$ in this example. A v-shaped probability curve is drawn separately for each previous crossover points, and an overall acceptance probability curve, denoted by the continuous line, is defined as the minimum of the individual probability curves. For the next chiasma location, a random point in the plane is selected. In the example, six such points are marked by a cross. If the random point is below the overall acceptance probability curve (as is the case for the three points labeled 'Accept'), the x value of that point is selected as the next crossover location. On the other hand, the presence of the point above the overall acceptance probability curve (like the three points labeled 'Reject') indicates that the intended location is too close to at least one of the previous crossover points. In such a case, the procedure is repeated starting from the selection of a new point in the plane.

the number of generations g, i.e., $O(ng)$. This is also the overall time requirement of the chromosome simulator.

If the number of individuals n in a generation grows exponentially, the time requirement is an exponential function of the number of generations. One should keep in mind that time consumption of this magnitude is unavoidable: the time requirement of merely printing the segments of the individuals in the last generation is $O(ng)$.

In practice, if the population size and age are not very large, the simulation is relatively rapid: simulating the inheritance of a single replicate of population BASIC (Figure 7.1) on an Ultra Sparc 10 computer takes approximately 55 seconds. The chromosome simulator has been implemented in C programming language.

4.4 Simulating disease phenotype and sample collection

In the next step, we consider the phase when the segment data of all individuals has been created by the chromosome simulator. The analyst probably wants to create a random sample of individuals from the final population, which may be too large to analyze as a whole. In addition, he/she may desire that a fixed number of individuals in the sample is affected by a (complex) disease, and another part is unaffected. Such a sampling strategy can be advantageous for association analysis of haplotypes, which we will return to in Section 5.1.2.

Keeping this in mind, we consider two interrelated problems: simulating the disease statuses and sampling affected and unaffected individuals from the population.

As input we have a pedigree containing data about the chromosomal segments for each individual occurring in it. Additionally, we require the genetic model (see Section 2.6) of the disease in question.

The way to determine the affection status of the individuals in the population depends on the representation of the disease model. Next, we consider three possible ways to describe the genetic model of a disease and introduce a method for determining the disease phenotype for each method. These three alternative ways of disease model representation and phenotype selection are implemented in the simulator software.

Method 1: specifying affected/phenocopy counts In this method, the number of carriers and phenocopies in each sample is fixed. The affected carriers are picked randomly from the group of individuals that carry the mutation and the affected phenocopies and control individuals are randomly selected from the group of individuals that do not have a mutation. There is an underlying assumption of a dominant effect, where at least one copy of the mutated allele is sufficient to cause the disease. The assumption that control individuals are required not to carry the mutation is equivalent with the assumption of a fully penetrant disease. This assumption could, however, be relaxed easily.

In the first method, the definition of disease phenotype and sampling are done by picking three types of individuals: affected mutation-carriers, affected non-carriers, and healthy non-carriers. The individuals in the first group carry at least one copy of the disease-predisposing mutation and, thus, have the disease phenotype. The members of the second group are phenocopies. They are individuals that do not carry the mutation but, still, have the disease phenotype. The disease of these individuals may be caused by other genes and/or environmental factors. The two groups of diseased individuals are generally indistinguishable to the researcher who is able to collect only a joint sample of affected individuals consisting of both affected carriers and phenocopies. By controlling the proportion of mutation carriers in the joint affected sample, it is possible to determine the genetic effect of the disease-predisposing mutation.

Method 2: use of penetrance matrices The disease model can be described using penetrance matrices and specifying the population frequencies of the disease alleles, as was described in [Ris90].

We can now define the genetic model of a disease to consist of two parts: a frequency distribution of the genotypes, as well as a penetrance of the disease with respect to each genotype. For simplicity, we assume the different disease-predisposing loci to be inherited independently. When this assumption holds, we can define a k-locus genotype as a tuple (P, Π), where P is a k-vector containing the probabilities of disease-predisposing alleles, and Π is a 3^k-hypercube of penetrances.

For example, the penetrances of a disease caused by either of two fully penetrant genes can be expressed by a 3×3 matrix, where the rows correspond to genotypes A_1A_1, A_1a_1, and a_1a_1, and the columns to genotypes A_2A_2, A_2a_2, and a_2a_2, respectively. For an example of a penetrance matrix, see Figure 7.7 on page 110.

The penetrance matrix representation of genetic disease can model the effect and interaction of any number of loci to the emergence of a genetic disease. The possible environmental effects are marginalized: the presence of a penetrance value other than zero or unity can be thought to represent the uncertainty caused by the environmental component.

Method 3: determining liability to the disease The third alternative to determine phenotypes is to compute for each individual his/her *liability* L to the disease. The individual's probability p of being affected depends on the liability via formula

$$\log \frac{p}{1-p} = L,$$

where probability p is a logistic transformation of liability L. Once the liability of an individual is determined, he/she is then labeled to be affected with probability p and unaffected with probability of $1-p$.

A complex liability-based disease model including several disease susceptibility genes and environmental covariates has been simulated for GAW12 workshop [WAA01]. In simulation of the GAW12 data, the liabilities were also used to create values of several quantitative traits. In our work, we select the liability-based approach for the experiments of chapters 6 and 8, and compute the liability of an individual from formula

$$L = \sum_{i=1}^{n_g} a_i x_{gi} + \sum_{j=1}^{n_e} b_j x_{ej} + c x_r + C.$$

Above, constant n_g refers to the number of genes and constant n_e to the number of environmental covariates. Constant a_i defines the emphasis of the effect of gene i, constant b_j that of environmental effect j, and constant c that of the random effect. Variable x_{gi} is an indicator for presence of a mutated allele in gene i[1], and variable x_{ej} the individual's value for environmental covariate j, which follows a standard normal distribution. Variable x_r, also following a standard normal distribution, corresponds to the individual's random component of the liability. Variable C is a baseline liability value that is used to obtain the desired prevalence of the disease. The value of C is determined in a special presampling phase. As a result of the procedure, the expected proportion of affected individuals in the population equals the desired population prevalence. The desired number of affected individuals can then be sampled from this set in linear time.

Collecting the affected samples The next task is to perform sampling. In case-control studies, two samples need to be collected: one that

[1] The user may define the number of different founder mutations in each locus. The different mutation alleles are treated equally in the model. Different founder mutations introduce allele heterogeneity to the samples.

contains affected individuals and one where the individuals are labeled as unaffected. In most linkage disequilibrium-based methods, the focus is then set on differences in the observed allele or haplotypes frequencies.

Since we aim at collecting random affected individuals in the population, the affected samples of desired size can be selected randomly from the individuals that were labeled as affected in the phenotype definition stage.

Collecting the control samples There are two major approaches to collecting the control sample: that of using random healthy individuals in the population, or using so-called *pseudocontrol*, or *family-based control chromosomes* , i.e., haplotypes constructed of nontransmitted parental segments (see Figure 4.6). Both chromosomes present at the affected offspring are labeled as affected. Next, two artificial chromosomes are constructed by merging the nontransmitted chromosomal segments of each parent. In the simulations where full segment data is available this can be done easily. However, in real disease mapping projects, the construction of pseudocontrol chromosomes needs to be based on marker data. In this case the *phase*, i.e., parental origin, of each allele is determined for each marker locus: The paternally inherited alleles in the consequent loci then constitute one haplotype and the maternally inherited alleles the other. Likewise, the remaining alleles in the parental genotypes constitute the control haplotypes. The determination of the phase can be done unequivocally except in the case where the offspring and the parents are identical heterozygotes in a marker locus. If we denote the single-locus genotypes of the parents as $\{p_{11}, p_{12}\}$ and $\{p_{21}, p_{22}\}$ and of the child as $\{c_1, c_2\}$, the haplotyping can trivially be done as indicated in Algorithm 12.

Algorithm 12 Haplotype determination based on single-locus genotypes of two parents $\{p_{11}, p_{12}\}$ and $\{p_{21}, p_{22}\}$ and one offspring $\{c_1, c_2\}$

if $\{p_{11}, p_{12}\} = \{p_{21}, p_{22}\} = \{c_1, c_2\}$ and $c_1 = c_2$ **then**
  give up
else if $c_1 = p_{11}$ and ($c_2 = p_{21}$ or $c_2 = p_{22}$) **then**
  report c_1 paternal, c_2 maternal
else if $c_2 = p_{11}$ and ($c_1 = p_{21}$ or $c_1 = p_{22}$) **then**
  report c_2 paternal, c_1 maternal
else if $c_1 = p_{12}$ **then**
  report c_1 paternal, c_2 maternal
else
  report c_2 paternal, c_1 maternal
end if

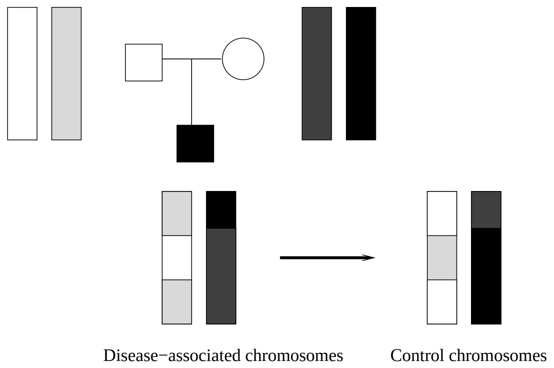

Figure 4.6: Sampling of disease-associated and pseudocontrol chromosomes. The black filling of the symbol for the offspring indicates that he is affected. From each trio (two parents and an affected offspring), the chromosomes of the offspring are taken as the disease-associated chromosomes, whereas the control chromosomes are constructed from nontransmitted segments of the parents. In practice, construction is done based on transmitted and nontransmitted alleles at marker loci.

The use of pseudocontrol chromosomes is considered favorable to the population-based controls, since the use of family-based control chromosomes may effectively alleviate the effect of population stratification.

The major drawback of population-based controls is that the affected individuals are expected to be more closely related than randomly picked, healthy individuals in a population isolate [TeW98, HVS99]. Thus, the use of population-based control individuals can lead to a situation, where the affected sample is expected to contain fewer haplotypes with higher frequencies. As a consequence, several haplotypes may show association to the phenotype, irrespective of the chromosomal region being studied.

An additional benefit of the pseudocontrol approach is the fact that the case and control individuals may share a more similar environment compared to cases and random healthy individuals, and the incorporation of family information also significantly eases the process of determining haplotypes.

Despite the benefits there is, however, a potential danger of selection bias in using pseudocontrol chromosomes. The bias is possible whenever the affected individuals with two participating parents differ genetically from individuals with one or no participating parents. Such a situation may occur if the disease is either associated to high mortality or reluctance to participate in the study. Bruun and Ewalds considered [BrE99] this as a potential problem for psychiatric disorders, where trios with healthy parents are likely to be overrepresented. From a genetic point of view, genes that have a dominant effect on the phenotype are consequently likely to be underrepresented and phenocopies overrepresented in the samples. [BrE99]

For the three methods of phenotype definition and sample collection (affected/phenocopy counts, penetrance matrices and liability) the time requirement is $O(n)$, where n is the number of individuals in the final population. In the implementation, multiple samples are collected in a single run of the sampler component. Collecting 100 samples by the penetrance matrix approach for population BASIC (Figure 7.1) on a Sun UltraSparc 10 takes approximately 125 seconds. The execution time for the liability approach is in the same ballpark. The space requirement is negligible; it does not depend on the number of individuals in the pedigree. All implementations were done in the C programming language.

4.5 Simulating generation of marker data

Once the segment data of sampled individuals is available, determining the genotypes is straightforward. The user describes a specification of a marker map, which contains the names, locations and allele frequencies of each marker.

As a consequence of the simulation procedure, true 'error-free' haplotypes are available, and they are also used in the experiments of this thesis. To reach maximum realism, the haplotyping procedure could alternatively be done based merely on information available from the genotypes in the trios. In each marker locus, the phases of the alleles cannot be unambiguously created, if all three members of the trio are identical heterozygotes. In all other cases, haplotyping is straightforward. The time requirement at this step is negligible.

4.6 Simulation-based estimation of significance and power

In this section we give two simple examples of the use of the chromosome simulator as an aid in a gene localization study, and briefly consider its potential use in the evaluation of the power of the intended study setup. We will return to the gene-mapping methods mentioned here later, in Section 5.1.

The two examples illustrate the use of the chromosome simulator in determining the statistical significance (empirical p value) of the observed positive gene-mapping result. The procedure for simulation-based estimation of empirical p value is outlined in [DSG96]. The idea is to repeatedly simulate the inheritance of the genotypes in the pedigree on the condition that no disease genes are present, i.e., the inheritance is purely Mendelian. In each iteration, the value of the test statistic is computed, and the observed value of the statistic (from the real data) is compared against that distribution. [DSG96]

In [HVS99], selected components of the simulator that we presented in this thesis (see Figure 4.7) were used to iteratively simulate inheritance in a large, extended schizophrenia pedigree under the null hypothesis of no association of the haplotype to the phenotype.

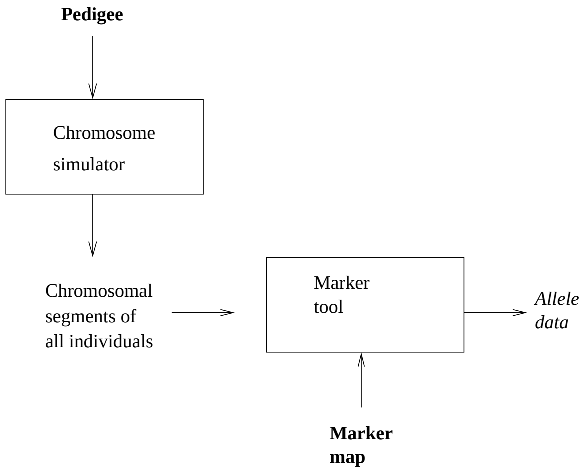

Figure 4.7: Components and data flow in the example of the schizophrenia [HVS99], where only selected components of the simulator software were used. The figure displays the input data (bold font), the output data (italic font) and the intermediate data of the software (regular font). The arrows indicate the flow of information between the user and the software as well as between the software components.

Among the observed findings was the fact that a single haplotype of four adjacent markers, spanning for the region of 6.6 cM, was shared identical-by-state by eight affected members of a single, extended pedigree. The test statistic was then defined as the length of the longest segment shared by at least eight affected pedigree members. The significance of the observation was first evaluated by treating all founder individuals in the pedigree as independent, and, based on this pedigree structure, creating artificial haplotype data. From these haplotype data, the distribution for the length of the longest chromosomal segment that was shared by at least eight affected pedigree members was computed, and it was observed that a shared region of 6.6 cM or more was observed only with probability 0.00029, indicating a significant finding. Because this kind of statistic can potentially give higher values, if there are hidden relationships in the pedigree, the procedure was repeated with three variants of the pedigree, where different degrees of kinship were introduced between the founders of the pedigree. The results of these modifications were, however, small. [HVS99]

A slightly more elaborate approach had to be used in a study on developmental dyslexia [NMH01], where the chromosome simulator presented in this thesis was used to obtain a genomewide significance level for the observation that in an extended pedigree 19 out of 21 individuals shared a haplotype of at least 16 cM. (The analysis of haplotype sharing (see e.g. [HBB94, VaM97]) will be further discussed in Section 5.1.2). The inheritance of genetic material in the pedigree was simulated one million times and it did not happen even once that 19 or more out of the 21 affected individuals in the pedigree shared a region as long or longer than the one observed. Even though that indicated that the p value was extremely close to zero, some additional calculations were needed to provide an upper bound for the p value.

Let us denote the probability that sharing as extreme as observed occurs in a single iteration by p_s. Consequently, such sharing is observed at least once in a million iterations with probability of $1-(1-p_s)^{10^6}$.

Next, let us assume that the p value was as high as $p_s = 10^{-5}$. For such a hypothetical p value of $p_s = 10^{-5}$, the probability of extreme sharing occurring at least once in a million iterations would be 0.999955, which indicates that the sharing would have been observed almost certainly at least once. Since that was not the case, it is safe to assume that for the true p value inequality $p_s < 10^{-5}$ holds with a very high probability. Finally, since the result was obtained from a 60.8 cM region, and there are approximately 55 such regions in the genome, a Bonferroni correction (see Section 5.1.4) was performed, and a new upper bound for p value was obtained: $p = 1-(1-p_s)^{55} < 1-(1-10^{-5})^{55} \approx 0.00055$, which led us to the conclusion that one can safely assume the genomewide corrected p value to be below 0.001.

The power of an intended study design can in principle be evaluated by entering the study-specific population history parameters, creating artificial data sets and analyzing them in the fashion intended. This is, however, dangerous, since the assumptions about the population history and the disease model may not reflect the real situation accurately. Thus, the results of evaluating the power of the intended study setup *a priori* should be considered suggestive at best.

Chapter 5

For maximal power at minimal cost

One of the goals of this work is to develop a computational methodology for the comparison of different approaches for disease gene localization. In this chapter we first review some of the commonly used methods for disease gene localization. In the latter part of this chapter, we introduce a framework for analyzing the cost of an intended study setup. This allows comparison of intended study designs, and it will later be used in the empirical analysis of linkage disequilibrium measures (Chapter 7).

5.1 Approaches to disease gene localization

We define the problem of disease gene localization, or gene mapping, to consist of two tasks: finding the genomic regions, where the susceptibility genes of the genetic disease are located with high probability, and providing adequate (statistical) evidence in favor of those regions.

The current gene mapping methods are based on analyzing genetic marker data. Each method is often considered to belong into either of the two groups: family-based *linkage methods*, where the target is to find inheritance patterns in extended pedigrees, or population-based *association methods* that aim at detecting alleles that are overrepresented in the affected sample. However, this categorization is not strict: there are methods that are able to combine some benefits of both the approaches.

In a genomewide gene-mapping project, a frequent order of the analysis stages is:

1. Linkage analysis with a sparse map. The target is to find prominent regions and exclude others.

2. Fine mapping using linkage methods. One tries to compensate for missing information in the data of the first stage by adding more

markers to the prominent regions that showed putative evidence of linkage in the first stage.

3. Association analysis. One tries to find alleles and haplotypes that are overrepresented among the affected members of the pedigrees. The aim is to narrow down the region showing positive linkage as much as possible. At this stage, the linkage information obtained from the pedigrees is often unusable, since the genetic distances are so small that only few, if any, recombinations can be observed in the pedigree data.

The data may be different for different stages. As the project proceeds, it may be necessary to obtain new marker genotypes for the individuals in the sample as well as collect new individuals to increase the sample size.

In the following sections, we review some existing linkage and association methods are reviewed, including some methods that combine ideas of the two approaches. Then we propose a method for finding the optimal cost for a genome scan project divided into multiple stages.

5.1.1 Linkage analysis

The purpose of linkage analysis (see e.g. [Mor55, KDR96, Ott99]) is to localize disease genes by detecting genetic linkage between disease-predisposing loci and the nearby marker loci.

Let us consider an ideal case of a monogenic disease, where there is a one-to-one correspondence between the phenotype and the existence of a mutated allele in the disease locus in each family (dominant mode of inheritance, see Section 2.6), and the marker map is infinitely dense and informative. In this ideal case, the closer the marker is to the disease locus, the more often the disease status and the marker alleles have been inherited together. Finally, when we look at a marker that is exactly on top of the disease locus, we can see complete correspondence between a marker allele and the phenotype. The marker allele (or alleles) involved may vary from family to family, but a similar pattern of joint inheritance can be seen in all families.

In reality, the marker maps are not infinitely dense and informative, and the true disease models can be complicated. The idea described above can still be used, but more sophisticated methodology is needed to handle the complications. The linkage methods are divided into two groups: *parametric* linkage analysis, where assumptions are made on the mode of inheritance of the disease, and *nonparametric*, or model-free, methods, where such assumptions are not made. According to [HBG96], one of the main advantages of parametric methods are their statistical efficiency, when the disease model is known, as well as their capability to provide an estimate of disease-gene location, and the fact that they permit both disease gene localization and exclusion of loci. The latter two restrictions are not necessarily limitations of nonparametric analysis, since the applicability of nonparametric analysis in disease gene localization was shown in [Ris90b], and an algorithm for disease gene localization using multiple markers simultaneously (*multipoint analysis*) was presented in [KDR96].

A review of the theory of parametric linkage analysis methods is found in [Ott99]. These methods are divided into two categories: *two-point* and *multipoint* methods. In two-point analysis, one marker is considered at a time, while in multipoint analysis information is combined from multiple adjacent markers.

The idea of performing sequential tests for linkage between two loci using LOD (logarithm-of-odds) scores dates back almost half a century and was first presented in [Mor55]. In two-point analysis the likelihood for the recombination fraction θ between the disease and marker loci is maximized. The likelihood of this estimate $\hat{\theta}$ is compared to the likelihood in case of no linkage between the loci ($\theta = \frac{1}{2}$), and the resulting quantity $Z(\hat{\theta}) = \frac{L(\hat{\theta})}{L(\frac{1}{2})}$ is called the maximum of a LOD score. By convention, and based on considerations of genomewide significance, a LOD score higher than 3.0 has been accepted as evidence of significant linkage [LaK95]. In most genetic studies, two-point LOD scores are still reported, but only as the first stage before multipoint analysis [Ott99]. For multipoint analysis of complex diseases thresholds of 2.2, 3.6, and 5.4 have been proposed [LaK95] for suggestive, significant, and highly significant linkage. These thresholds are expected to be exceeded 1, 0.05, and 0.001 times in a full genome scan by chance only [LaK95].

Parametric linkage analysis is often conducted by Linkage [LLJ84] and GENEHUNTER [KDR96] computer programs. The latter can also be used to perform nonparametric analysis. In nonparametric linkage analysis, the focus is on affected sibling pairs (ASP), or more generally, affected relative pairs (ARP). The methods are based on computing identical-by-descent statuses of the marker loci and comparing the distribution to the null hypothesis under no genetic linkage. The ASP method is computationally simple: based on observed inheritance from parents to offspring, one first computes, how many alleles (0, 1, or 2) each pair of affected siblings shares identical-by-descent. If the observed numbers of pairs sharing 0, 1, and 2 alleles are n_0, n_1, and n_2, then the expected numbers under null hypothesis are $N/4$, $N/2$, and $N/4$, consecutively, when $N = n_0 + n_1 + n_2$. Now, the deviation from the expected values can be tested by a chi-square test. The underlying idea is that significant deviations are a consequence of selective sampling: since the sample contains an excess of affected individuals, some of them share a common disease allele (which may vary between families), and, as a consequence, have excess IBD sharing in chromosome regions around the disease locus. In the areas that do not contain the disease loci, the individuals should not share more alleles in the marker loci than would be expected by chance only.

The other method, ARP, is a generalization of the ASP method to affective relative pairs of arbitrary kinship. The framework for nonparametric linkage analysis, as implemented in the GENEHUNTER computer program [KDR96], comprises of computing probability distribution for different inheritance vectors conditional on the genotypes at the marker loci, computing the value of the scoring function, which is the likelihood ratio, i.e. the probability of the phenotype given the inheritance vector divided by the sum of probabilities for the phenotype over all inheritance vectors, normalizing the score, and, finally, combining the scores obtained from different families [KDR96].

Due to the shift of interest towards multifactorial diseases, the performance of parametric multilocus linkage analysis was analyzed in [MSK93]. The results were, however, discouraging. First, the determination of a disease model requires specifying a large number of parameter values, which hardly can be done accurately. Second, the possible gain of using a multilocus linkage statistic seems very limited: most of the information incorporated in the multilocus linkage statistic is eventually seen in the sum of the corresponding two-point statistics. Thus, the method suggested in [MSK93] is to combine individual correlated two-point linkage statistics, but that procedure is unfortunately subject to statistical difficulties. [MSK93]

To end the review of parametric and nonparametric linkage methods, it should be noted that this classification into two groups may after all be nonwarranted. Even though the analyst does not have to explicitly specify the disease model in nonparametric methods, the methods may still contain built-in assumptions about the underlying model. For instance, it has been demonstrated [KSB94] that certain ‘nonparametric’ sib-pair tests are equivalent to parametric linkage analysis with a recessive model.

5.1.2 Association and linkage disequilibrium analysis

The terms ‘linkage disequilibrium analysis’ and ‘association analysis’, or mapping, are used as synonyms in this thesis, since a distinction between the two concepts is rarely made.

Sometimes, the term 'LD mapping' refers specifically to fine-scale mapping, where a model of linkage disequilibrium between nearby loci is applied (e.g. [Guo97, HDK92, XiJ99]), and the task is to find the disease locus by detecting common inheritance of the disease locus and the flanking marker alleles to the affected individuals. The (genomewide) association studies, on the other hand, can be thought to focus on the observed association between the disease phenotype and individual marker loci without considering the underlying model of LD. In both cases, the phenomenon to detect (association of the marker allele and the disease locus caused by LD between the two loci) is the same, which makes a sharp distinction between the concepts unwarranted.

Association (or LD) analysis uses information about the inheritance of chromosomal segments. Let us consider what happens to a chromosome in the course of repeated meioses. At the top of Figure 5.1 a single founder chromosome is shown. In the middle of the chromosome, there is a locus containing a disease-predisposing allele. In the ideal case of a one-to-one correspondence between the mutation and the disease, the affected individuals still share the common mutation when the generations pass, but the length of the conserved segment around the mutation becomes shorter. Thus, the members of the current population share not only the common disease allele, but also a chromosomal segment derived identical-by-descent from one of the founder individuals. In this region, the individual has inherited his/her marker alleles ($(1, 2, 1)$ in the example) directly from the founder chromosome. As a consequence, affected individuals in the sample share not only the common mutation, but also common marker alleles around the disease locus. This is the phenomenon that is used in disease gene localization by association analysis.

Relative risk approaches One of the first efforts to formalize the concept of association analysis dates back to 1955, with the introduction of the relative risk (RR) approach [Woo55] for measuring the association between blood groups — used as markers — and different diseases. The test measures whether the carrier of one variant has a higher risk of being affected than carriers of another variant.

A distinction can be made between association methods that assume genotype data and those that require haplotype data (where the haplotypes can be either true or hypothetical). The haplotype relative risk [RWC81, FaR87] is a family-based association test that uses genotypes constructed from nontransmitted parental alleles as control genotypes, which reduces environmental heterogeneity between cases and controls. The

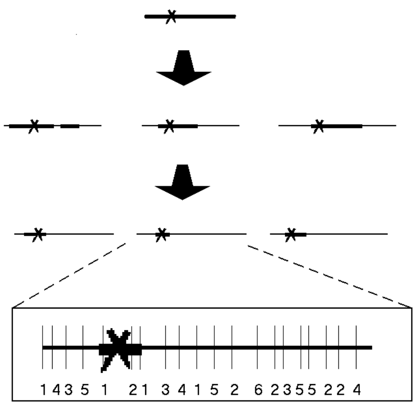

Figure 5.1: Decay of linkage disequilibrium, a marker map and a conserved haplotype. At the top of the figure, a single founder haplotype is illustrated by a thick line, and its allele in the disease locus is marked by 'X'. When generations pass, the chromosomal segment inherited IBD from the founder chromosome diminishes. In the final population (the bottom three chromosomes), all individuals that have inherited the common 'X' allele in the disease locus have also inherited a segment of the founder chromosome around that locus. At the bottom of the figure, one of the current chromosomes is magnified, and it can be seen that together with the allele at the disease locus the individual has inherited all marker alleles (1,2,1) that are in the region of the conserved segment.

method was further improved by Terwilliger and Ott, who introduced the so-called haplotype-based haplotype relative risk (HHRR) approach in 1992 [TeO92]. In an HHRR test for an arbitrary allele, a 2×2 table is created:

Allele	H	$\bar{H}$	Total
Disease-associated (transmitted)	W	X	N
Control (not transmitted)	Y	Z	N
Total	W+Y	X+Z	2N

The HHRR test statistic is now a test for the null hypothesis of E(W)=E(Y), which is a standard chi-square test of independence on a 2×2

table, and can be expressed as:

$$\frac{2N(WZ - XY)^2}{(W + X)(W + Y)(X + Z)(Y + Z)}.$$

If all marginal frequencies are kept unchanged, then the expected frequencies under the null hypothesis of no association are:

Allele	H	$\bar{H}$	Total
Disease-associated (transmitted)	(W+Y)/2	(X+Z)/2	N
Control (not transmitted)	(W+Y)/2	(X+Z)/2	N
Total	W+Y	X+Z	2N

If the observed frequencies in the former table are denoted by o_{ij} and the expected frequencies in the latter table by e_{ij}, where $i, j \in \{1, 2\}$, the chi-square test statistic is then computed from expression

$$\sum_{i=1}^{2} \sum_{j=1}^{2} \frac{(o_{ij} - e_{ij})^2}{e_{ij}}.$$

Under the null hypothesis, and assuming fixed marginal frequencies, the test statistic follows asymptotically a chi-square distribution with one degree of freedom. The test statistic described above is designed for diallelic loci, and there are alternative ways to generalize it into a multiallelic case (see [Ter95]). If there are m alleles, the simplest way is to perform a single statistical test on a $2 \times m$ contingency table, but, unfortunately, the degrees of freedom increase, and, additionally, some of the observed frequencies may be small enough to render the χ^2 test unreliable. An alternative approach is to perform m different statistical tests and use the value of the highest score as a test statistic. The latter approach requires correction for multiple testing (see section 5.1.4).

TDT methods Transmission/disequilibrium tests (TDT, [SGE93]) are among the most widely used family-based association methods. They test excess transmission of alleles from heterozygous parents to affected offspring, and are capable of detecting within-family linkage in the presence of linkage disequilibrium between the marker and disease loci. Among the strongest advantages of the TDT methods are their simplicity and robustness to population stratification. As an example, consider a collection of nuclear families (two parents and an offspring), where n parents are heterozygous at a fixed marker locus with respect to allele a. If the marker being studied is unlinked to the disease gene, there has been no bias in selection of the families that would have favored carriers of any particular allele to be overrepresented among the affected sample members. Thus, since Mendelian inheritance is assumed, it is equally likely that either of the alleles of a heterozygous parent has been passed to the offspring. Next, consider the case where the locus is linked to the gene. Now, it is likely that there is overrepresentation of families, where the disease-associated marker allele, and, consequently, the disease, has been passed to the offspring. This overrepresentation is seen as transmission distortion, and it is the phenomenon that TDT detects.

In case of an unlinked locus, suppose that allele a was transmitted n_1 times, and the other allele n_2 times, where $n_1 + n_2 = n$. Under the null hypothesis of unlinked loci, the random variable X indicating the prior probability for the number of times allele a was transmitted follows a binomial distribution: $X \sim Bin(n, \frac{1}{2})$. The observed count n_1 can simply be compared against this distribution, or, alternatively, a chi-squared test can be performed. The statistic is computed from formula

$$\chi^2 = \frac{(n_1 - n_2)^2}{(n_1 + n_2)}$$

and it follows asymptotically a chi-squared distribution with one degree of freedom, when the null hypothesis holds.

Maximum likelihood methods Terwilliger ([Ter95]) proposed a likelihood-based method that is extendable to multiple alleles and multiple adjacent markers. In the model, there is one parameter λ expressing the degree of association of an allele to the disease phenotype. If the disease-associated allele is known, the log-likelihood of the data is first computed for each allele in a marker locus as a function of λ and then summed over all alleles, weighting by the allele frequencies. Since the associated allele is generally not known a priori, all alleles are considered as disease-associated in turn, and a posterior probability distribution for the associated allele is obtained.

The approach is extended to multiple loci by combining the likelihoods over adjacent markers to form a composite likelihood. Even though the likelihood incorporates pairwise degree-of-association information from multiple markers, it does not utilize the fact that the association information over multiple markers is obtained from continuous, shared haplotypes among the affected individuals, a fact pointed out by [SLF99]. Also, there is a general problem [Ter95] that the results of linkage disequilibrium-based methods may be hard to interpret because of possible violations of model assumptions that may be due to events in early recombination history as well as possible errors in intermarker distances.

The ancestral haplotype reconstruction (AHR, [SLF99]) method is designed for genome screening purposes, and, according to [SLF99], it should be applicable to mapping of complex traits in presence of allele and locus heterogeneity. (To be specific, in the experiments of [SLF99], genetic heterogeneity was introduced by increasing the probability of carriers of a normal genotype to be affected. Such a method corresponds to locus heterogeneity, where the affection status of the non-carriers of the risk allele in one locus is explained by other unlinked loci.)

In the AHR method, the length of the shared haplotype is included in the analysis. The method estimates four parameters: the disease locus, the proportion of disease-associated chromosomes that are descended from the common founder chromosome, the separation time from the common ancestor, and the frequencies of the marker alleles. Both AHR and Terwilliger's likelihood-based approach perform well in the absence of locus heterogeneity and when the marker density is high enough [SLF99]. In the presence of locus heterogeneity and with sparse maps, both methods perform poorly, but in the intermediate situations AHR is demonstrated to perform better. In the power simulations, full linkage equilibrium was assumed between adjacent markers in the control population, which eliminates all background linkage disequilibrium and may create an unrealistically optimistic view of the performance. However, the isolated false-positive results caused by background linkage disequilibrium need not be a problem, since the method is designed for the initial stage of a genome scan. In the presence of locus heterogeneity, however, AHR seems superior to methods that look at loci independently, and the results clearly suggest that LD-based mapping methods utilizing multiple markers simultaneously can be used in the first stage of a genome scan even for complex diseases. However, an unknown degree of this superiority can probably be explained by the fact that AHR uses haplotype data, and in the comparison of [SLF99] it could utilize the true, error-free simulated haplotypes whereas Terwilliger's likelihood-based method [Ter95] uses genotype data that inherently contains less information.

The decay of haplotype sharing method (DHS, [McS99]) is most suitable to situations where very dense marker maps are available. DHS is a likelihood-based method and it accommodates missing data and marker mutations and can also cope with situations when haplotype data can be obtained only partially.

Haplotype sharing methods Recall from Figure 5.1 that carriers of a disease-predisposing allele of common origin have also inherited a region of the founder chromosome IBD from the common ancestor. Then, in an ideal case of a simple monogenic recessive disease, all chromosomes of the affected individuals carry the common disease allele and a variable region around the disease locus. The segments shared IBD around the disease locus are expected not only to be present but also to be longer than the average length of a shared chromosomal segment, as was demonstrated both analytically and empirically in [TSL97]. The haplotype sharing methods are designed to detect such presence of the common shared haplotypes of (possibly) excess length around the disease locus.

The idea of haplotype sharing was first successfully utilized in [HBB94] for localization of benign recurrent intrahepatic cholestasis, where the observation was that five (out of six) chromosomes of three distantly related affected individuals shared a common region of about 20 cM. The calculations of [HBB94] suggested that such sharing was extremely unlikely to occur by chance only, which could then be taken as the proof of gene localization.

The haplotype sharing statistic [VaM97] summarizes the sharing of haplotypes around a locus into a one-dimensional test statistic. In the method, all haplotypes are compared pairwise, the length of the region shared IBS (overlapping a fixed locus) is computed, and the standard deviation of these lengths is used as the haplotype sharing statistic. The use of standard deviation is rationalized in [VaM97] by the observation that excess sharing generally manifests itself in a few long shared haplotypes accompanied with many whose lengths are close to zero.

The maximum identity length contrast (MILC, [BGH01]) statistic is computed from pedigree data. The method compares all haplotypes in the data pairwise and calculates for each marker the mean length of the region shared identical-by-state. The marker-wise differences between the mean haplotype identity length of transmitted and nontransmitted haplotypes are then computed, and the maximum of these differences among all markers is then taken as the MILC statistic. The p value of the observation is obtained using a bootstrap procedure. It is pointed out in [BGH01] that the method is feasible for study of complex diseases using information from pedigrees of arbitrary size.

Bayesian, data mining, and other methods Within Bayesian paradigm, Morris et al. [MWB02] have recently proposed a Markov-Chain Monte Carlo method for obtaining the posterior probability density function for the location of the disease gene. Beginning from the observed data, the algorithm generates the 'shattered coalescence history' which allows several disease-predisposing founder mutations (allele heterogeneity) in the same locus. Sporadic cases that do not link into the shattered genealogy are also allowed. These sporadic cases are thought to be caused by other disease-predisposing loci (locus heterogeneity) or environmental factors. The method is applicable for fine-scale mapping [MWB02].

Haplotype Pattern Mining (HPM, [TOV00]) is one of the first attempts to utilize data mining approaches in disease gene localization. In HPM, frequent haplotype patterns are sought from case/control data, their degree of association to the phenotype is tested, and markers are scored based on the number of overlapping associated haplotypes.

An example of the function of HPM is shown in Figure 5.2. The algorithm looks for haplotype patterns that are sequences of consecutive markers that are allowed to contain gaps, which are marked by an asterisk in the figure. The algorithms could be implemented very efficiently using the data mining methods [AMS96, MaT97]. For each pattern, the association to the phenotype is measured by a traditional chi-square test computed from a 2×2 contingency table, and all patterns exceeding a given threshold are included in calculation of marker-wise scores. A score for each marker is calculated using a simple scoring function: it is the number of associated patterns overlapping the marker. The advantages of haplotype pattern mining include its ability to efficiently combine information from adjacent markers and multiple founder haplotypes as well as its capability to deal with missing and erroneous data. The data sets used in the power calculations of the article [TOV00] were generated using the population simulator software described in Chapter 4.

Recently, Schaid et al. published a collection of score tests applicable in association mapping that can also be used in cases where haplotyping is ambiguous [SRT02]. The computation of the statistics is rapid, but, like HPM, the empirical p values for the observed association need to be obtained by a permutation procedure (see page 65). The use of generalized linear models, however, allows environmental covariates and traits of various forms to be incorporated easily [SRT02].

Computational aspects The computational complexity of association methods depends on the number of haplotypes evaluated. Some methods

A. Most strongly disease-associated patterns

Marker	01	02	03	04	05	06	07	08
Pattern01	4	1	6	6	3	*	*	*
Pattern02	4	1	6	6	3	6	*	*
Pattern03	4	1	6	6	*	6	*	*
Pattern04	4	1	6	*	3	6	*	*
Pattern05	4	1	*	6	3	*	*	*
Pattern06	4	1	*	6	3	6	*	*
Pattern07	4	*	6	6	3	*	*	*
Pattern08	4	*	6	6	3	6	*	*
Pattern09	4	*	6	*	3	6	*	*
Pattern10	*	1	6	6	*	*	*	*
Pattern11	*	1	6	6	3	*	*	*
Frequency	9	11	11	11	10	6	0	0

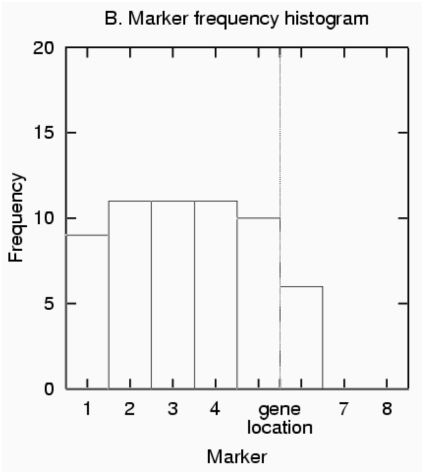

Figure 5.2: Idea of haplotype pattern mining [TOV00]. (A) a collection of strongly associated patterns. The 11 haplotypes have been chosen because of their strong association to the phenotype, measured by a χ^2 statistic. There are 8 markers in the example, and an asterisk in a pattern indicates that any allele is allowed in that locus. For each marker, the frequency of overlapping patterns is calculated. Gaps are also counted. (B) Marker-wise scores displayed graphically. The true gene location is indicated by a vertical line. HPM predicts the gene to be located at the marker with the highest score.

(e.g. [SLF99]) evaluate the likelihood of all possible haplotypes that can be created from the allele data. As a consequence, the computation time is an exponential function of the number of markers. In the easiest case of m biallelic markers, 2^m evaluations are required, which in practice restricts the applicability of the methods to short haplotypes of a few (5 or less) markers only. In any case the computational time for multiallelic markers increases even more rapidly.

The computational time can, however, be efficiently reduced by restricting oneself to methods that evaluate only those haplotypes that actually appear in the data instead of all potential haplotypes created as combinations of individual alleles in consequent loci.

If the haplotype length is fixed, it is obvious that the maximum number of different haplotypes to be evaluated is twice the number of individuals in the data. With this restriction, the computational time grows as a linear function of the number of markers.

In case the length of the haplotypes is not fixed, like in data mining-based algorithms (such as HPM [TOV00] and extension of HRR [RWC81, FaR87] to multilocus haplotypes of arbitrary length), the time requirement is still tolerable. It can be seen that for the set of n chromosomes and haplotype length of at most m markers, the maximum number of haplotypes appearing in the data is only $\frac{m(m+1)n}{2}$. This follows from the observation that the startpoint i of each haplotype can be chosen from set $\{1, 2, \ldots, m\}$ and the endpoint j from set $\{i, i+1, \ldots, m\}$. For the region spanning markers i to j, there are at most n different haplotypes.

5.1.3 Combining linkage and association

Recall that gene mapping methods have traditionally been classified into two groups: linkage methods (Section 5.1.1) that utilize recombinations within known pedigrees and association methods that take advantage of the hidden recombination history in the population (Section 5.1.2). Now, the question of the relationship between the two approaches arises. Can there be one without another, and, if yes, why? This question is addressed in [GrD96], where a distinction is made between 'necessary' loci and 'susceptibility' loci. The former are prerequisites for emergence of a disease, and should be observed in both linkage and association analysis. The latter are, however, alleles that increase susceptibility to the disease but are neither necessary nor sufficient for the disease, and may remain undetected in linkage analysis. According to [GrD96], a test of heterogeneity can be used to determine the type of the source of linkage disequilibrium: in case of the susceptibility model, the associated allele is likely to segregate with the disease in some families, whereas the other alleles show no patterns of linkage. On the other hand, in case of a 'necessary' locus and the linkage disequilibrium model, no heterogeneity should appear.

The pedigree disequilibrium test (PDT, [MMW00]) combines information from nuclear families and discordant sibships (pairs of siblings, where one is affected and the other unaffected) in general pedigrees. In the method, a pedigree-specific measure for linkage disequilibrium is defined as

$$D = \frac{1}{n_T + n_D}\left(\sum_{j=1}^{n_T} X_{Tj} + \sum_{j=1}^{n_S} X_{Sj}\right)$$

where n_T is the total number of trios, n_S the total number of discordant sib pairs, X_{Tj} the number of times allele j was transmitted in trio n_T minus the number of times it was not transmitted, and X_{Sj} the number of times the allele was present in the affected member of the discordant sibship minus the times it was not. The linkage disequilibrium measures D_i obtained from individual pedigrees are then summarized to produce a joint statistic of LD written as

$$T = \frac{\sum_{i=1}^{N} D_i}{\sqrt{\sum_{i=1}^{N} D_i^2}}$$

where N is the total number of informative pedigrees. The statistic asymptotically follows a standard normal distribution under the null hypothesis of no linkage disequilibrium, and can thus readily be used as a test statistic. According to [MMW00], PDT offers substantial advantage to traditional methods that only use a subset of the data by selecting a single trio from each extended pedigree. The method is also robust to population substructure and can be efficiently used also in cases where extended pedigree information is not available.

5.1.4 Multiple testing problem and its solutions

Multiple testing is present in almost every gene mapping study. The problem emerges from the fact that when several hypotheses are tested using a predetermined nominal significance level, the probability that some — or at least one — of the tests produces a significant nominal p value increases as a function of tests. In gene mapping, typical sources for multiple testing include:

- testing different alleles in a marker locus for association;
- testing different markers for association;
- testing different haplotypes — or patterns — of adjacent markers for association;
- repeating parametric linkage analysis with another disease model;
- changing the phenotype criterion and repeating the linkage or association analysis;
- repeating the analysis with additional markers and/or individuals in the sample.

If all of the tests performed can be considered independent, one can apply the Bonferroni correction: if the nominal error level is α and there are n independent tests performed, the overall nominal p value is then $1-(1-\alpha)^n$, which for $\alpha n \ll 1$ is about αn.

One factor is common to all sources of the multiple testing problem listed above: the individual tests are not independent or even close to it. Thus, applying the Bonferroni correction results in grossly elevated p values, which makes the correction practically useless.

There are other problems with Bonferroni correction — and with correction to multiple testing in general: the fact that interpretation of one test depends on the number of other (possibly independent) tests done, the inevident increase in type II errors, and the problematic selection of which tests should be included [Per98]. Due to disputable questions in selecting the proper way of correction in linkage studies, it has even been suggested [Els97] that the authors only publish nominal p values, which can then be properly interpreted by the reader.

Also, there are cases when significant p values are not the aim of the analysis, but they are merely used in ranking individual markers, as is done in the first stage of the two-stage designs proposed in [SVV02].

If correction for multiple testing is desired, there are less conservative ways to do it than Bonferroni correction. In case/control setups, a Monte Carlo permutation test [ChD94, Goo00] is an efficient way to correct for multiple testing. Permutation tests are nonparametric statistical tests that can be used in situations where the observations are exchangable and the two groups of observations (here: affected and control chromosomes) are taken from the same distribution under the null hypothesis [Goo00]. The data collected for association analysis meet these criteria: the measures of association do not depend on the order of chromosomes in the data, and both the affected and the control haplotypes can well be assumed to be random samples from the general population of chromosomes, if the haplotypes are completely independent of the disease statuses.

The Monte Carlo permutation test procedure is outlined as follows:

$k := 0$;
Compute statistic t from real data d;
for i:=1 **to** N **do begin**
    Create permutated data d_p, where phenotypes are shuffled;
    Compute statistic t_p from permutated data d_p;
    if $t_p > t$ **then**
        $k := k + 1$;
end;
p:=k/N;

In the permutation procedure, the null distribution is created by repeatedly shuffling the phenotypes. This approach is suitable to a case/control setup, where there are no close relationships between the individuals, allowing the chromosomes to be considered independent. In the presence of family data, a possible approach suggested in [DSG96] is to simulate the inheritance of genotypes in the pedigrees conditioning on the known genotypes of the unaffected individuals.

In the permutation procedure described above, the shuffling of dichotomous phenotypes can easily be done in linear time. The permutation procedure correctly takes into account the interdependencies between individual markers and alleles, since no assumptions are made of the independence of the tests.

In the ideal case, one should compare the observed data to all possible permutations, but this is practically impossible and is approximated by random shuffling of phenotypes. The lower the p values are, the larger the required number of iterations and the computational time demands are. This may make the permutation procedure an unattractive alternative, and in some real-life cases days or even weeks may be needed to obtain corrected p values.

5.2 Power and cost of disease gene localization

Disease gene localization often begins from a situation where there is little knowledge about the true locations of the disease-predisposing genes. In such a case, all regions of the genome are probably considered equally likely candidate regions, and a full *genome scan* is carried out.

A genome scan can be divided into multiple *stages*. In transition to the next stage, the aim is to exclude regions that are unlikely to contain any of the disease-predisposing variants. Thus, for each point of the region, a decision rule is applied, and as a consequence, each point is classified into one of the two groups, i.e., those qualifying for the next stage and those that are excluded. As a result of the procedure, the overall region to be considered diminishes to a fraction of what it was at the previous stage. Before transition to the next stage, additional samples may be collected and additional genotypes created. In reality, the first stage of a genome scan is often linkage analysis, which is then followed by linkage disequilibrium analysis.

Multistage designs are in practice applied in most genomewide studies. Satagopan et al. have recently conducted experiments on two-stage designs in gene mapping [SVV02], and they suggest a general guideline that 75%

of the recources of a study should be allocated to the first stage, after which the remaining budget should be allocated for analyzing the most prominent 10% of markers. Next, we propose a method for finding an optimal strategy for disease gene localization that, unlike [SVV02] takes into account the effects of population history, and optimizes over different measures of linkage disequilibrium.

5.2.1 Stages, power, and false positives

Figure 5.3 shows a genome scan consisting of four stages. The stages are numbered from 0 to 3. Stage 0 corresponds to the initial situation, where no analysis has been done, and the whole genome is considered as a candidate region. At this stage, markers may be placed throughout the genome using a relatively sparse map, and, as a result of computer analysis, transition to stage 1 occurs. At stage 1, regions that are unlikely to contain the disease gene have been excluded, and a denser map of markers is used for the remaining chromosomal segments. The resulting allele data is then analyzed, and the promising regions have been further narrowed, when stage 2 takes place. At stage 2, the remaining chromosomal segments are sequenced, its genes identified, and the more accurate locus of the disease-causing mutation estimated. The last transition leading to stage 3 corresponds to finding the exact location of the gene.

Gene localization strategies can be compared by analyzing their costs. The more effectively the exclusion of unlikely regions is performed, the lower the overall cost of the study will be. The cost can be decreased by excluding more regions, but, on the other hand, this increases the probability that the true disease-causing mutation is not among those genomic regions that survive to the final stage. In this chapter, we present a cost model that allows comparison of genome scan strategies, provided that relative costs of the proposed methods can be stated.

Consider stage i of a genome scan, and denote by A_i the finite set of continuous, non-overlapping chromosomal segments that are available in stage i of a genome scan. Set A_0 always refers to the entire genome. The *power* of stage i to detect the disease-predisposing locus d can be expressed in terms of probability as $F_i = P(d \in A_i | d \in A_{i-1})$. The corresponding *false positive rate* p_i gives the total size (as a fraction of the size of the region in the previous stage) of the genetic region that qualifies for the next stage provided that no disease-causing mutation is present, that is $p_i = \frac{|A_i|}{|A_{i-1}|}$, given that $d \notin A_0$. The false-positive rate can be computed from simulated data sets, where no disease gene is present.

Alternatively, the *qualification rate* q_i of the stage can be computed.

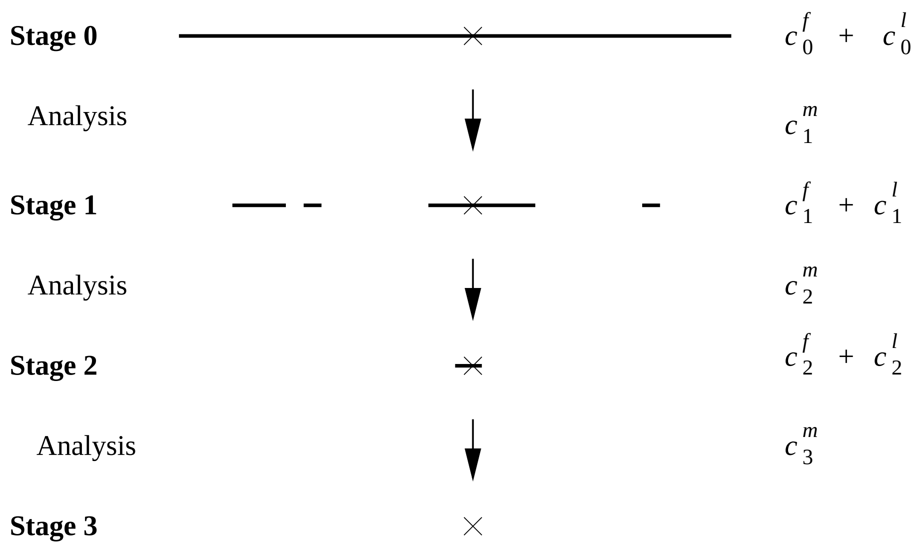

Figure 5.3: A genome scan consisting of four stages and the corresponding cost components. The horizontal line at each stage shows the part of the genome that remains to be analyzed. The (hidden) disease locus is indicated by 'X'. The fixed and length-specific costs of each stage i are denoted by c_i^f and c_i^l. The cost for missing the locus in moving to stage i is indicated by c_i^m.

It represents the proportion of the region that qualifies for the next stage, provided that the true disease locus does that. The total region qualified, thus, contains the region around the true disease locus, and possibly other regions that can be considered as by-products of false positive signals. The qualification rate can be expressed as $q_i = \frac{|A_i|}{|A_{i-1}|}$ given that $d \in A_i$, and it can be computed from the same data sets that were used in the assessment of power, i.e. those computed under the hypothesis of a disease gene present. In the initial stage of a genome scan, where the size of the region showing a positive signal is small compared to the entire size of the region, the false positive rate p_i can be approximated by using the corresponding qualification rate q_i.

5.2.2 Obtaining the power function

Each stage in a genome scan involves a tradeoff between power and cost. Clearly, power F_i is a monotonic increasing function of the respective proportion q_i of genome qualifying from the previous stage, and also of the false-positive rate p_i. Now, two important questions arise: what criteria should be used in selecting chromosomal regions to reach desired power, and how can one compute the power function? These questions can be answered with the help of the population simulator. Let us concentrate on a region that contains $K+1$ markers and K intervals. For simplicity, we assume that the mutation locus is searched at granularity of an intermarker interval. The disease locus is assumed to be located *a priori* in any point of the chromosome with equal probability. Thus, the longer the segment, the more likely it contains the mutation. Let us denote the genetic lengths of the K intervals by l_i, where $i \in \{1, \dots, K\}$. In Figure 5.4, a simple case of $K=3$ intervals is shown.

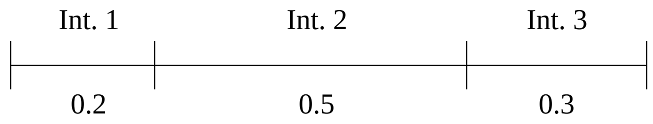

Figure 5.4: An example of a chromosomal region with four markers (denoted by vertical lines) and three intermarker intervals (denoted by 'Int. 1', 'Int. 2', and 'Int. 3') with relative lengths of $l_1 = 0.2$, $l_2 = 0.5$, and $l_3 = 0.3$. The relative lengths should sum to unity and be proportional to the genetic lengths of the intervals.

In this context, the term *localization method* refers to any procedure or function that takes as an input the marker map specification and the genotype (or haplotype) data and produces a numeric output value for each intermarker interval. The desired property of a good localization method is that the output value (here referred to as the *test statistic value*) for an interval should have a monotonous relationship to the probability of the disease gene being at that interval.

To empirically analyze the behavior of a method, we use the chromosome simulator (see Section 4.3) to produce artificial data sets that are created under two alternative hypotheses. The distributions of the test statistics are then compared to find out the relationship between the power and the false-positive rate. The less the distributions overlap, the better the statistical performance.

First, it is necessary to find out how the test statistic is distributed for each of these intervals, when the disease-causing variant (disease allele) is present by chance only. To do so, N artificial data sets are created with a 'null' disease model. For each data set and for each interval, the value of the desired test statistic T is computed. Let us denote a single test statistic value obtained in iteration r for interval i by $T^+_{i,r}$.

Next, one needs to find out how the statistic is distributed, when the disease is present in the data. Thus, N data sets are now created under the alternative hypothesis. Since there are K intervals altogether, this results in NK test statistic values. Based on these values, K different empirical distributions of the test statistic are obtained, one for each interval. A test statistic value obtained in iteration r for interval i is here denoted by $T^D_{i,r}$, where $r \in \{1, \ldots, N\}$.

Now, the interval-specific false positive rates corresponding to a given power level f can be computed. To do this, the following procedure is repeated for each interval $i \in \{1, \ldots, K\}$:

1. Find a threshold t_i such that a fraction f of the test statistic values $T^D_{i,r}$, $r \in \{1, \ldots, N\}$ computed under the alternative hypothesis are equal to or exceed that threshold.

2. Compute the fraction of observations $T^+_{i,r}$, $r \in \{1, \ldots, N\}$ created under the null hypothesis that are equal to or exceed threshold t_i. Denote this probability by $p(i, f)$.

Thus, for any interval i and power level f, a false-positive rate $p(i, f)$ is obtained. Consecutively, the power f can be written as a function of interval-specific false-positive rates.

Consider the simple example of Figure 5.4. Let us assume that the following statistic values were obtained for the three intervals:

Int.	Null hypothesis	Alternative hypothesis
1	5, 3, 2, 7, 4, 6, 2, 2, 1, 3	7, 8, 6, 7, 9, 8, 7, 10, 7, 6
2	6, 5, 6, 3, 5, 4, 6, 3, 5, 7	5, 6, 6, 7, 8, 7, 6, 4, 7, 8
3	2, 1, 3, 1, 2, 3 ,2, 3, 1, 2	5, 4, 5, 4, 6, 5, 4, 5, 2, 4

Here, the thresholds for the power level of, for example, 0.8, are $t_1 = 7$ $t_2 = 6$, and $t_3 = 4$. The respective interval-specific false positive rates are $p(1, 0.8) = 0.1$, $p(2, 0.8) = 0.4$, and $p(3, 0.8) = 0$.

Finally, since the disease susceptibility locus is expected to be anywhere in the region with equal probability, the overall false-positive rate $P(f)$ related to a power level f can be obtained by weighing the false-positive rate within each interval with the length of the interval:

$$P(f) = \sum_{i=1}^{K} p(i,f) l_i.$$

Thus, the resulting false-positive rate can be written as the function of power, and vice versa.

In the example, the overall false-positive rate related to the power level of 0.8 is $P(0.8) = 0.2 \cdot 0.1 + 0.5 \cdot 0.4 + 0.3 \cdot 0 = 0.22$. This is the expected total length of segments that do not contain the disease-causing mutation but still qualify for the next round, if an overall power of 0.8 is required, and the segment-specific thresholds of $t_1 = 7$, $t_2 = 6$, and $t_3 = 4$ are used as selection criteria.

This overall power vs. false-positive function can be used in performing cost calculations. Finally, when the desired false-positive rate and the respective overall power level have been obtained, the corresponding threshold value of the statistic for each interval can be computed.

5.2.3 Cost components and minimization

At each stage j of the genome scan the cost can be divided into three components that were also depicted in Figure 5.3. These components are:

1. fixed costs (c_j^f);
2. length-specific costs (c_j^l);
3. cost of risk for missing the disease locus (c^m).

The fixed cost includes costs that do not depend on the length of the genomic area that is to be scanned. This includes things such as sampling costs. Different methods may require different sample sizes and sample strategies yielding different fixed costs. The length-specific cost caused by the stage depends linearly on the size of the region present in the stage. The longer the region to be considered, the more markers are needed and the higher the cost will be. Finally, the risk of not being able to detect the correct disease-predisposing locus should be expressed in terms of expenses. For this purpose, constant c^m denotes the fixed cost caused by the negative result that the analyst faces in the unfortunate case where the existing gene will eventually not be found in the project.

The total cost C of an $n+1$-stage genome scan (consisting of stages $0, \ldots, n$) can now be expressed as a function of the cost components, which in turn are written as functions of the false-positive rates at each stage and cost constants π_i^l, π_i^f, and π^m given by the analyst:

$$C = \sum_{j=0}^{n-1}(c_j^f + c_j^l) + \sum_{j=1}^{n} c_j^m \tag{5.1}$$

where

$$c_j^l = \pi_j^l \prod_{k=0}^{j} P_k,$$
$$c_j^f = \pi_j^f,$$

and

$$\sum_{j=1}^{n} c_j^m = \pi^m(1 - \prod_{j=1}^{n} F_j(P_j)).$$

The aim is to find such a combination of tolerated false-positive rates $(P_1, \dots, P_n)$ that minimizes the whole expression.

In addition, the overall budget of the study may be limited to some fixed value. The fixed costs π_i^f as well as the length-dependent costs π_i^l for different stages of the analysis can be evaluated in terms of expenses more or less accurately. In finding out the length-dependent cost constants the key question to ask is "What proportion of the overall budget would it take, if the whole genome was scanned with this method?"

However, the cost for a false negative result π^m is harder to find out, and the value used will be based on the judgements of the researcher, and factors such as his/her willingness to take risk. The researcher may be able to estimate the risk directly, or, alternatively, he/she might approach this problem by asking the question: "What is the lowest tolerable power level for the study, and what is the maximum allowed size of the potential chromosomal segments related to it, given that the overall budget of B cannot be exceeded?". Denote these two quantities by F^* and P^*. Next, one might consider how much extra the researcher would be willing to invest, if this extra investment resulted in 100% power and a target region of equal size P^*. Denote this subjective extra investment by kB. Now, the theoretical cost of a failure (π^m) can be computed from equation

$$B + (1 - F^*)\pi^m = (1 + k)B,$$

where the left side corresponds to the intended budget plus the cost of uncertainty caused by the fact that the true locus may be missed. This is set equal to the subjective tolerable cost of a study with full power, allowing us to compute the cost of failure as

$$\pi^m = \frac{kB}{1 - F^*}.$$

This coarse approximation allows, in principle, evaluating the cost of failure on the same scale as other costs at different genome scan stages.

Having obtained the cost constants and the simulated power functions F_i, the overall cost of a method can be written as a function of the false-positive rates P_i solely. This cost is then minimized and the most cost-effective method accompanied by the optimal power level can be found by looking at the global minimum among the family of cost functions, each corresponding to a different approach.

5.2.4 Example of cost computation

As an example, let us consider an imaginary genome scan that is intended to consist of the following steps:

Step	Task	Cost
Stage 0	Collect samples make genotypes, sparse map for the whole genome	c_0^f c_0^l
Transition $0 \to 1$	LD analysis	c_1^m
Stage 1	Collect additional samples, make genotypes, dense map for targeted regions	c_1^f c_1^l
Transition $1 \to 2$	LD analysis	c_2^m
Stage 2	Sequencing	c_2^f c_2^l
Transition $2 \to 3$		c_3^m
Stage 3	Final location	

The first task is to find a value for the subjective cost π^m of missing the true disease locus. In this case, a budget of 1 is allocated to the genome scan project, and the ultimate aim of the decision-maker is to reach power of 85 per cent and finish up at a region of 3 centiMorgans. However, the decision-maker would be willing to invest an extra cost of 0.75, if he/she obtained complete certainty of finding the true locus. The subjective cost of a failure π_m can now be approximated to be $\pi^m = 0.75/(1 - 0.85) = 5$.

Next, the other constants for costs are determined. For stage 0, the fixed cost constant is set to $\pi_0^f = 0.1$ and the length-dependent cost to $\pi_0^l = 0.12$.

Since stage 0 is the first stage in the study, the length specific cost is applicable to the entire genome. The genotypes are analyzed computationally during transition to stage 1, and, as the result of the analysis, the chromosomal regions are ranked with respect to their probability of containing any of the disease-predisposing variants.

The power vs. false-positive curves for transitions between stages can be obtained with the population simulator the way described in the previous section. Let us assume that in our example the simulated curves turns out to be those of Figure 5.5.

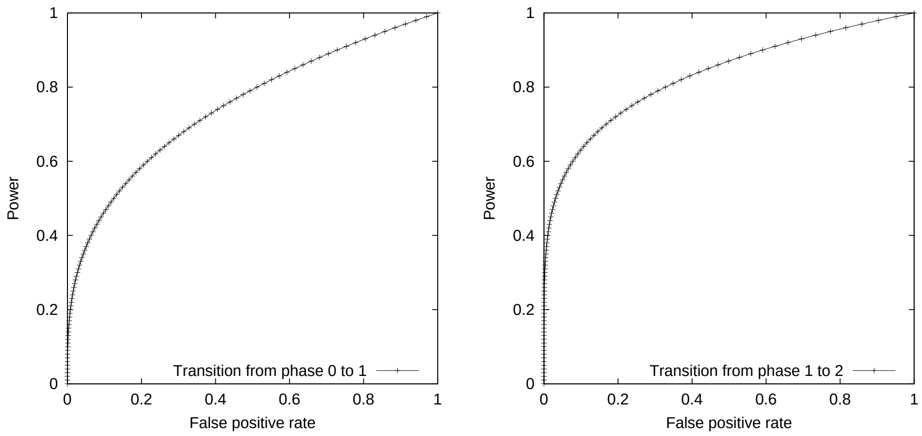

Figure 5.5: An example of power vs. false positive rate curves for transitions between stages 0 and 1 (on the left) and stages 1 and 2 (on the right). The false-positive rate is on the horizontal axis and the corresponding power on the vertical axis.

Next, it is assumed that the fixed cost constants for stages 1 and 2 are the same as for stage 0, i.e., $\pi_1^f = \pi_2^f = 0.1$. The length-specific costs are set to $\pi_1^l = 2$ for stage 1 and $\pi_2^l = 20$ for stage 2. Furthermore, we assume that the final localization does not contain any risk for missing the gene, i.e. $F(P_3) = 1$.

Now, the overall cost related to the genome scan, as defined in equation 5.1, can be written as (note that $P_0 = 1$):

$$\begin{aligned} C &= \sum_{i=0}^{2}(\pi_i^f + \pi_i^l \prod_{j=0}^{i} P_j) + \pi^m(1 - \prod_{i=1}^{3} F(P_i)) \\ &= 0.1 + 0.12 \cdot 1 + 0.1 + 2.0P_1 + 0.1 + 20.0P_1P_2 + 5(1 - F(P_1)F(P_2)) \\ &= 5.42 + 2.0P_1 + 20.0P_1P_2 - 5F(P_1)F(P_2). \end{aligned}$$

The cost is now a function of two unknown parameters, P_1 and P_2, and can be plotted as a three-dimensional surface (Figure 5.6).

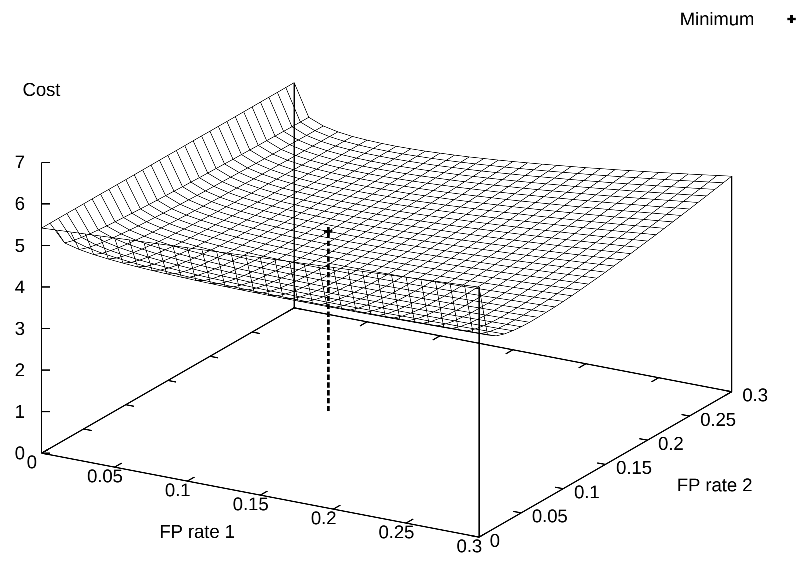

Figure 5.6: An example of a cost surface. The axes labeled as 'FP rate 1' and 'FP rate 2' contain the false-positive rates for transitions from stage 0 to 1, and 1 to 2, respectively. The vertical axis corresponds to the cost caused by the two false-positive rates. The minimum of the surface (denoted by '+') corresponds to the optimal combination of false-positive rates that yields the lowest cost.

In the example, the minimum cost is 4.33, which is obtained by selecting false-positive rates of $P_1 = 0.11$ and $P_2 = 0.15$, yielding power levels of 0.48 and 0.68 between stages and an overall power level as low as 0.33. The low power level here only reflects the fact that π^m (the cost constant for missing the gene) was relatively low.

Comparison of minimum costs obtained by different methods is a way to compare their performance relative to each other. This approach will be used for linkage disequilibrium mapping methods in Chapter 7. For each method, the empirical power vs. false positive curves are computed, and the methods are compared by applying selected sets of cost constants considered appropriate for genome scans.

Chapter 6

Population history and conserved chromosomal segments

Linkage disequilibrium mapping relies on the assumption that there has been only one — or at most a small number — of initial founder mutations. Each of the founder mutations has been associated to an initial *founder haplotype*, originally covering the entire chromosome. When the time and generations pass, this founder haplotype is split into smaller and smaller segments by repeated recombinations. In this chapter we present experiment results on the length and number of conserved and shared haplotypes in isolated populations that are generated using the population simulator. Some gene mapping methods that utilize haplotype sharing (e.g. [VaM97, BGH01]) were reviewed in Section 5.1.2.

Recall that linkage disequilibrium is defined as nonrandom association between two or more loci. It can be measured, e.g., by the pairwise linkage disequilibrium coefficient $\delta = p(AB) - p(A)p(B)$, where p(A) and p(B) denote the population frequencies of allele A in one locus and allele B in another locus [TeW98]. In this section, we analyze the amount of linkage disequilibrium between loci. To achieve this, we use the population simulator to analyze the nature of conserved segments of ancestral chromosomes in the collected population samples. Such analysis cannot be based directly on marker data, since IBD information cannot necessarily be deduced from markers. However, the selected modular approach of first simulating inheritance of chromosomal segments (see Section 4.1) provides us with full IBD data for the simulated populations. Such data become very useful in the experiments of this chapter.

The special emphasis here is on potential uses of sharing of segments in haplotype association methods. Specifically, there are three questions to answer:

1. What is the extent of IBD sharing in randomly picked chromosomes, and how large segments of founder haplotypes do individuals in a random sample share IBD by chance alone?

2. What is the extent of IBD sharing in pseudocontrol chromosomes (see page 45), collected from parents of affected individuals? Does the degree of sharing depend on the population history?

3. What is the extent of IBD sharing in disease-associated chromosomes collected from affected sample members? Is there observable excess of sharing of chromosomal regions in the genome in general? How are these phenomena affected by population history?

The extent of IBD sharing is measured by a collection of summary sharing statistics that we created for this purpose.

6.1 The experiment design

We conduct simulation experiments to analyze the degree of IBD sharing. In the simulations we use a set of population history descriptions with variation in initial population size, spatial substructure, and number of generations. The differences between the 13 population history descriptions are summarized in Table 6.1. For all populations, the size of the final population was 100,000, and the internal migration rate between two adjacent subpopulations was 0.01. This indicates the probability that an individual born in one of the subpopulations migrates to an adjacent subpopulation before having children. There was no immigration from outside the entire population.

The definition of the 13 population histories allows analyzing the effect of population substructure (populations 1, 2, 3, 7, 4, and 5), number of generations (populations 10, 7, 11, 12, and 13), and number of founder individuals (populations 6, 7, and 8), while the other parameters stay constant.

The data sets for each population history and disease model are simulated in six steps;

1. Create a population pedigree. Populations are simulated generation by generation, maintaining the desired division into the subpopulations as well as internal immigration. The process is described in detail in Section 4.2.2.

2. Inherit chromosomal segments (see Section 4.3).

Label	Init.size	Generations	Structure
1	200	20	1×1
2	200	20	1×2
3	200	20	1×3
4	200	20	1×5
5	200	20	3×3
6	100	20	2×2
7	200	20	2×2
8	500	20	2×2
9	200	20	1×1
10	200	15	2×2
11	200	25	2×2
12	200	30	2×2
13	200	50	2×2

Table 6.1: Population history descriptions used in the analysis of observed linkage disequilibrium. The populations were simulated using distinct generations (see Section 4.2.2). Each of the 13 population history descriptions consists of the initial population size, the number of generations since foundation and the pattern of population substructure. For instance, expression 2×2 indicates that the population consists of four interconnected subpopulations (in a form of 2×2 grid). The migration probability for all 13 populations was 0.01, and the final population size 100,000.

3. Select a random mutation locus and the desired number of random mutation sources , i.e., founder chromosomes, for each sample. For the 'null model' with no disease gene, no mutations were selected, whereas for the model with strong genetic component, four different founder mutations in the same locus are selected. The presence of four disease mutations causes moderate allelic heterogeneity in the samples.

4. Compute the liability for each individual, and, based on the liability, decide, who is affected. For details of computation of the liability, see Section 4.4. The data sets where no disease was present were created using a disease model, where the coefficient of the genetic component was zero. For the model with a strong genetic component (corresponding to one disease gene with allelic heterogeneity), the expression for liability was $L = 5x_g + 3x_r + C$. In this expression, variable x_g is a binary indicator for the individual carrying *any* of the four founder mutations, whereas variable x_r represents a random component that follows a standard normal distribution in the population. Variable C

is a baseline liability that is used to fix the prevalence of the disease to the value of 0.05, which corresponds to a typical prevalence of a common multifactorial disease.

5. Randomly collect 200 affected individuals and their parents to each sample.

6. For members of each sample, create parental pseudocontrol chromosomes (see Section 4.4).

As a result of the simulation procedure, the data consists of segments of the chromosomes of the affected individuals (created both under the strong disease model and null model) as well as those of nontransmitted parental chromosomes.

From the data of the offspring (created either under the disease model or null model) and their parents, we define three *chromosome classes*. First, we analyzed pseudocontrol chromosomes of random individuals (marked as Class 1 in Figure 6.1). They are constructed from the nontransmitted chromosomes of affected individuals created under the null model. They represent completely random chromosomes sampled in the population. Second, we analyzed pseudocontrol chromosomes of affected individuals (Class 2 in Figure 6.1). The difference to the first class of chromosomes was that a true disease model was used instead of a null model. These correspond to 'normal' family-based control chromosomes that would be collected in a population-based association study. Third, chromosomes of affected individuals were collected (Class 3 in Figure 6.1). A true disease model was used, and the segments of the affected offspring were collected instead of the pseudocontrol chromosomes. The chromosomes created this way correspond to 'disease-associated' chromosomes that would be obtained from affected individuals in an association study.

In the experiments, each population (one large pedigree) was generated once for each set of parameters and each disease model (true model and null model), and 100 data sets were independently sampled from the final generations of each large pedigree. A higher number of repetitions would have provided more accurate estimates, but when some of the runs were repeated (data not shown), the results were consistent, indicating that the number of iterations was probably adequate. This also supports the assumption that a single creation of the population pedigree is sufficient to produce accurate results.

For each chromosome class, we create a collection of 100 data sets. In each data set, the sample consists of 200 affected offspring and their parents.

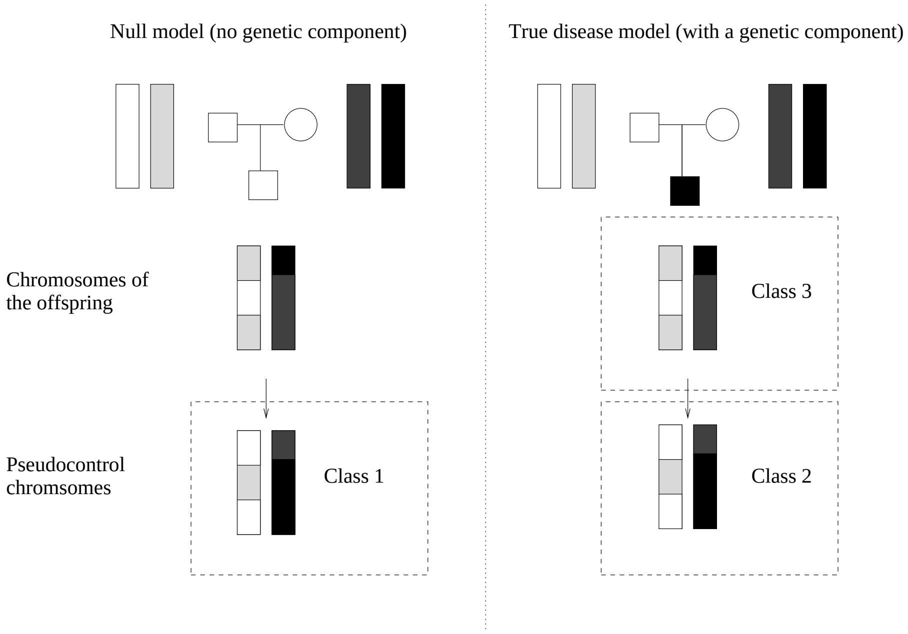

Figure 6.1: Construction of the three classes of chromosomes analyzed. The data generation was done under two disease models: a null model with no genetic component and a true disease model, where the presence of any disease allele had a considerable effect on the disease phenotype. For both disease models, pseudocontrol chromosomes were constructed from the nontransmitted chromosomal segments of the offspring in the samples. The pseudocontrol chromosomes generated under the null model constitute the first class of chromosomes. The second and third class are generated using the true disease model. The former is comprised of pseudocontrol chromosomes and the latter of the disease-associated chromosomes of the affected offspring.

This results in 400 chromosomes for each sample. The genetic length of the simulated region was set to 100 cM.

For each data set, we compute a collection of summary sharing statistic values, all indicating the degree of sharing ancestral chromosomal segments IBD. The statistics that we designed for this purpose, are:

NUMCHR number of such different founder chromosomes that overlap a fixed locus and are present in at least one copy in the sample. No sharing of the segment between the sample members is required.

AVGLEN average length of the chromosomal segment that overlaps a fixed locus, i.e., the average length of a conserved founder segment. No sharing of the segment between the sample members is required.

MAXLEN maximum length of an ancestral chromosome segment overlapping a fixed locus, computed over all chromosomes in the sample. No sharing of the segment between the sample members is required.

MAXINH maximum number of chromosomes where a single ancestral chromosome was observed (shared between the sample chromosomes) in a fixed locus.

EXn number of chromosomes, where any chromosomal segment of length greater than n (in centiMorgans, $n \in \{5, 10, \ldots, 30\}$) was present in a fixed locus. No sharing of the segment between the sample members is required.

INHn number of such ancestral chromosome segments overlapping a fixed locus that were shared between at least n chromosomes ($n \in 2, 3, 5, 10$).

The values of these sharing statistics are computed in locations 15, 25, ..., 85 cM, yielding eight different values for each statistic in each data set. An example of this is given in Figure 6.2, corresponding to statistic AVGLEN and population model 4 (Table 6.1, page 79). As is the case in this example, there is little variation in the values between chromosomal locations and individual data sets, and this variation appears to be completely random.

6.2 Results and conclusions

For each of the 13 population models and three chromosome classes, the averages and standard deviations of each sharing statistic (over the 8 locations

```
# Statistic: AVGLEN
# File:      stats.pop.4b.4.dis.M0

D#/Loc    0.15    0.25    0.35    0.45    0.55    0.65    0.75    0.85          avg     stdev
-------------------------------------------------------------------------------------------------
   1 :  0.0973  0.1042  0.1047  0.1110  0.1098  0.0996  0.0999  0.0971 :    0.1029    0.0050
   2 :  0.1017  0.1040  0.1004  0.1018  0.1066  0.0985  0.1022  0.0951 :    0.1013    0.0033
   3 :  0.1042  0.1132  0.1095  0.1109  0.1075  0.1036  0.1024  0.0961 :    0.1059    0.0051
   4 :  0.1060  0.1070  0.1080  0.1110  0.1136  0.1051  0.1010  0.0950 :    0.1058    0.0054
   5 :  0.1037  0.1055  0.0995  0.1074  0.1056  0.1011  0.0977  0.0960 :    0.1021    0.0039
   6 :  0.1046  0.1053  0.0989  0.1011  0.0987  0.0971  0.1021  0.1008 :    0.1011    0.0027
   7 :  0.0955  0.1029  0.1022  0.1093  0.1039  0.1043  0.1005  0.0926 :    0.1014    0.0049
   8 :  0.1052  0.1099  0.1140  0.1077  0.1028  0.1016  0.0992  0.0954 :    0.1045    0.0056
   9 :  0.1051  0.1087  0.1059  0.1090  0.1052  0.1053  0.1030  0.1002 :    0.1053    0.0027
  10 :  0.1097  0.1111  0.1015  0.1097  0.1097  0.1060  0.0982  0.0974 :    0.1054    0.0053
  11 :  0.1037  0.1057  0.1026  0.1045  0.1019  0.0951  0.0979  0.0992 :    0.1013    0.0034
  12 :  0.1039  0.1054  0.1015  0.1108  0.1081  0.0995  0.1025  0.0972 :    0.1036    0.0042
  13 :  0.1044  0.1103  0.1121  0.1110  0.1056  0.1055  0.1054  0.0960 :    0.1063    0.0048
  14 :  0.0973  0.1057  0.1110  0.1061  0.1028  0.0960  0.1080  0.1048 :    0.1040    0.0048
  15 :  0.1024  0.1058  0.1063  0.1017  0.1069  0.0991  0.1017  0.0984 :    0.1028    0.0030
  16 :  0.1010  0.1076  0.0981  0.1049  0.1051  0.0982  0.1009  0.0952 :    0.1014    0.0039
  17 :  0.0988  0.1037  0.1069  0.1094  0.1124  0.1032  0.1033  0.1021 :    0.1050    0.0040
  18 :  0.0983  0.1050  0.0985  0.1071  0.1054  0.1020  0.1003  0.0954 :    0.1015    0.0038
  19 :  0.1017  0.1048  0.1000  0.1036  0.1040  0.0956  0.0980  0.1017 :    0.1012    0.0030
  20 :  0.1047  0.1080  0.1137  0.1074  0.1063  0.0976  0.1067  0.1005 :    0.1056    0.0046

(omitted)

  98 :  0.0996  0.1087  0.1090  0.1037  0.1119  0.1027  0.1029  0.1009 :    0.1049    0.0041
  99 :  0.1015  0.1091  0.1014  0.1092  0.1080  0.1060  0.1004  0.0979 :    0.1042    0.0041
 100 :  0.1006  0.1043  0.1010  0.1021  0.1063  0.1024  0.0991  0.0938 :    0.1012    0.0035
-------------------------------------------------------------------------------------------------
avg     0.1008  0.1058  0.1042  0.1066  0.1062  0.1007  0.1021  0.0977 :    0.1030    0.0047
stdev   0.0032  0.0034  0.0041  0.0037  0.0034  0.0038  0.0036  0.0032 :=====================
```

Figure 6.2: An example of the output on values of sharing statistic AVGLEN, computed in locations $15, 25, \ldots, 85$ cM for selected data sets (1-20, 98-100). The averages ('avg') and standard deviations ('stdev') are shown for the 8 different locations of each data set (on the right), and 100 different data sets in each location (in the bottom). The overall average (0.1030) and standard deviation values (0.0047) are underlined. The table corresponds to chromosome class 1 (pseudocontrol chromosomes for random individuals).

and 100 data sets) are summarized for different population models in Appendix A. In computation of these values, the statistic values are combined from different data sets and chromosomal locations. The tables indicate that the differences in the sharing statistic values between the three classes are negligible, whether the chromosomes are pseudocontrol chromosomes collected from parents of random individuals (chromosome class 1), those collected from parents of affected individuals (chromosome class 2), or the affected individuals themselves (chromosome class 3).

The effect of population substructure is shown in Figure 6.4. The figure indicates that the differences between populations of different structure are small, assuming the moderate migration rate of 0.01. The result suggests that intrapopulation migration of this magnitude might be safely ignored in association-based studies. The effect of the number of generations, is, however, considerable, as indicated in Figure 6.5, and so is that of the number of initial founders (Figure 6.6). The figures show the sharing statistic values for chromosome class 1 (pseudocontrol chromosomes of random individuals); the values for the other two classes are highly similar as indicated by the tables of Appendix A.

First, let us consider how the observations obtained empirically by the population simulator correspond to the theoretical results that can be, under some assumptions, obtained on lengths of conserved chromosomal segments. To answer this question, we compare some of the obtained values to those that were analytically derived in [McS99]. In [McS99], the occurrence of crossovers in a chromatid is considered as a stationary renewal process, with mean interarrival time of one, corresponding to the genetic definition of Morgan. Based on [McS99], let us derive the expected length of chromosomal segment around a fixed point, marked by the vertical line in Figure 6.3. In each generation, we are interested in the length of the chromosomal segment inherited identical-by-descent from the ancestral chromosome. This segment is marked by the thick grey line in the figure.

For simplicity, let us consider Haldane's no interference model, where the interarrival density is defined as e^{-x}, where x represents the genetic distance in Morgans. Assuming non-interference eases the calculations, since, as a consequence, the length of the conserved segments both to the right and to the left of the variant are independent.

After the first meiosis, the probability that a crossover does not occur within length a to the right from the origin is equal to

$$1 - \int_0^a e^{-x}\, dx = e^{-a},$$

and the probability that this does not occur in g generations is equal to e^{-ag}. Consequently, the cumulative distribution function for the location of the crossover closest to the variant location in g generations is expressed as $F(a) = 1 - e^{-ag}$. The corresponding distribution function of a can be written as $F'(a) = ge^{-ag}$, which is a density of a random variably following exponential distribution with parameter $1/g$. Thus, since the lengths to the left (denoted by random variable X_1) and right (denoted by X_2) of the origin are equal due to assumed non-interference, the expected length of the overall segment is $E(X_1 + X_2) = E(X_1) + E(X_2) = 2E(X_1) = 2/g$. For instance, for a population of 20 generations, the expected value of the inherited segment around a variant is 10 cM. The average values of observed segments (AVGLEN) obtained from simulated data closely correspond to the theoretical values.

Since variables X_1 and X_2 are independent exponentially distributed

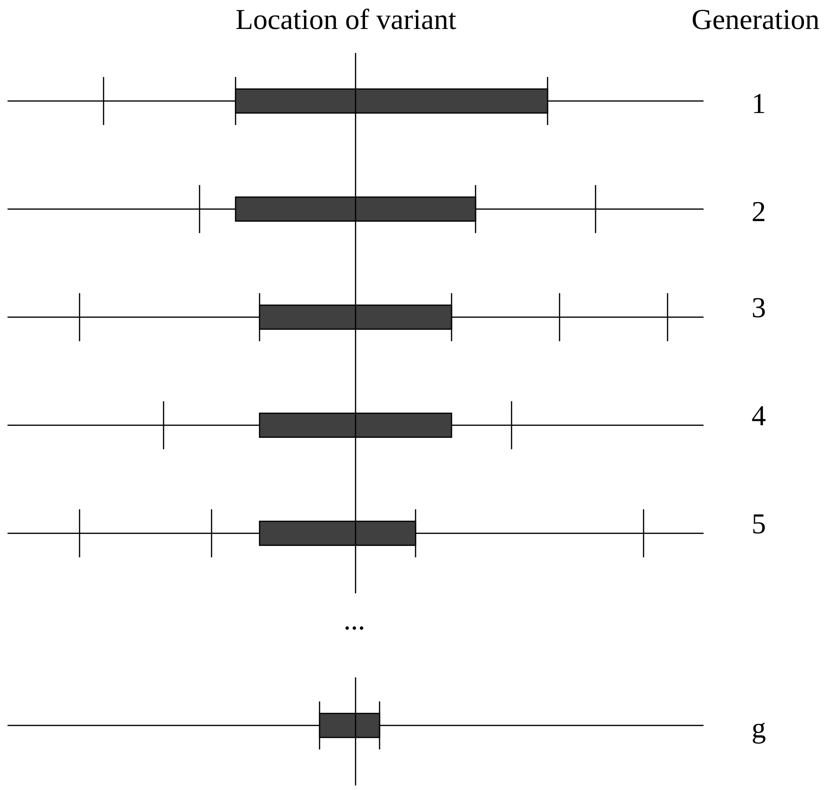

Figure 6.3: An example of the effect of repeated meioses on a chromosomal segment around a fixed variant. It is assumed that an individual in generation g has inherited a fixed variant from the ancestral chromosome in generation 0. In each generation, the length of the conserved segment may decrease. This happens if the crossover closest to the variant location, separately to the left and to the right of the variant, is closer than any of such locations in the previous generations. The conserved segment is denoted by a thick grey line, and the crossover locations are denoted by vertical bars.

Length (cM)	Theoretical value	Population								
		1	2	3	4	5	6	7	8	9
5	294.3	295.4	293.9	297.0	298.2	297.4	295.6	296.6	295.0	300.8
10	162.4	161.4	163.8	167.1	168.6	167.7	166.7	167.0	162.3	171.1
15	79.7	78.3	80.7	82.7	86.4	85.4	83.1	84.2	80.6	87.0
20	36.6	33.8	36.3	37.6	39.3	38.7	38.4	38.7	35.9	38.5
25	16.2	14.5	15.7	15.9	17.8	17.1	17.7	17.2	14.6	17.2
30	6.9	6.2	6.5	6.5	7.8	6.9	7.5	7.6	5.7	7.4

Table 6.2: Theoretical (based on [McS99]) and observed numbers of chromosomes, where the given length of segment is exceeded. The empirical results are computed from 400 nontransmitted pseudocontrol chromosomes that are obtained from random samples of individuals in the nine populations (numbered from 1 to 9, see Table 6.1) with a population history of 20 generations. The observed numbers are averaged over 100 data sets and 8 locations in each set.

random variables, their sum, which is the total length of the segment around the variant, follows a gamma distribution:

$$X_1 + X_2 \sim Gamma(2, g^{-1}).$$

This result is compared to values of sharing statistics EX5, EX10, EX15, EX20, EX25, and EX30, which show the frequencies of inherited chromosomal segments exceeding predefined lengths. Again, the EXn statistics correspond well to the theoretical calculations of [McS99], as Table 6.2 indicates.

The number of ancestral chromosomes observed in a fixed location in each sample (NUMCHR) is around 100 in all populations irrespective of the chromosome class, meaning that – due to genetic drift – in any selected locus three quarters of the original chromosomes have become extinct. Even though population substructure could have caused even higher genetic drift, and, consequently, segments from fewer founder chromosomes in the final generation, this was not the case with the population models analyzed. This probably indicates that the subpopulations have been large enough and the migration rate sufficiently high for strong genetic drift not to occur.

Since the length of conserved segments depends solely on number of meioses, it is not surprising that values for populations of similar age do not differ. Since the number of founder variants present in the final generation (NUMCHR) is almost constant with respect to the population model, it appears logical that the number of chromosomes inherited to a specified number of individuals (INH2, INH3, INH5, INH10) is also almost constant. This, again, might indicate that genetic drift has not been strong enough to create fluctuations in allele frequencies in any of the populations.

As Figure 6.5 shows, the number of generations since foundation of the population isolate affects the degree of sharing identical-by-descent (IBD) in the population. Due to the higher degree of genetic drift present in older populations — where the population expansion has been slower — the MAXINH sharing statistic indicating the frequency of the most common founder chromosome increases with the number of generations, while statistic NUMCHR (indicating the number of different IBD origins at a given locus) decreases. Other statistics related to lengths of the shared segments decrease as functions of the number of generations, as was expected. Again, considerable differences between the three chromosome classes (random parental chromosomes, parental chromosomes of affected individuals, and chromosomes of affected individuals) cannot be seen, which is probably an indication of the fact that the selected statistics summarize the sharing over the entire 100 cM region, but are unable to capture the slightly increased sharing around the disease locus in the group of disease-associated chromosomes (chromosome class 3).

The effect of the number of founders is depicted in Figure 6.6. The sharing statistics behave as expected in absence of disturbing factors: since the number of generations stays constant, no differences are observed in statistics derived from the lengths of the shared segments.

To conclude, the implemented simulation procedure allowed analysis of segments shared identical-by-descent in populations of various histories. The number of generations turned out to be the only factor in population history among those analyzed that had a clear effect on lengths of segments shared by members of the current population. Also, as expected, the number of initial founders is reflected in the observed frequencies of shared segments: the more founders, the less sharing. However, even though it is known that population substructure can lead to fluctuations in the frequencies of individual alleles [ThN97], this seems not to be reflected in the length of the shared segments in the general population, at least not with the selected migration probability. Assuming a complex disease model like the one used, the differences among the three chromosome classes (chromosomes of random individuals, parental nontransmitted chromosomes of affected individuals, and chromosomes of affected individuals) are not visible in the averages and variances presented in the results section.

Finally it should be mentioned that the results of this section define the limits for the applicability of association-based methods: the excess sharing among disease-associated chromosomes, although definitely hard to detect by chromosome-level aggregate measures, can find a manifestation in associated haplotypes, a phenomenon that is captured by methods analyzed in Sections 7 and 8.

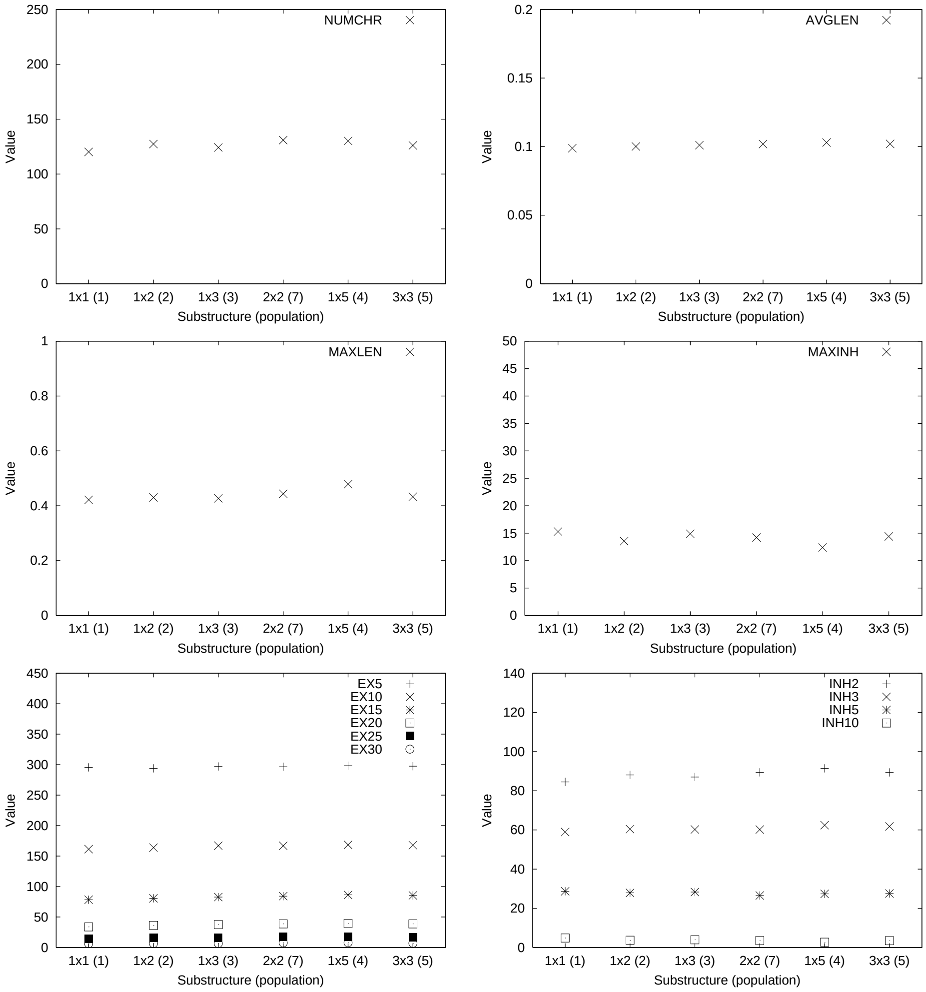

Figure 6.4: The effect of population substructure on sharing statistics measuring the extent of conserved chromosomal segments in chromosome class 1 (pseudocontrol chromosomes of random individuals). The populations (substructure type and population number) are indicated on the horizontal axis. The vertical axis show the values of the statistics NUMCHR and AVGLEN (at the top), MAXLEN and MAXINH (in the middle), and EXn and INHn (at the bottom). For substructure type, the notation 1×3, for example, indicates that the population consists of 3 subpopulations in the shape of a 1×3 grid.

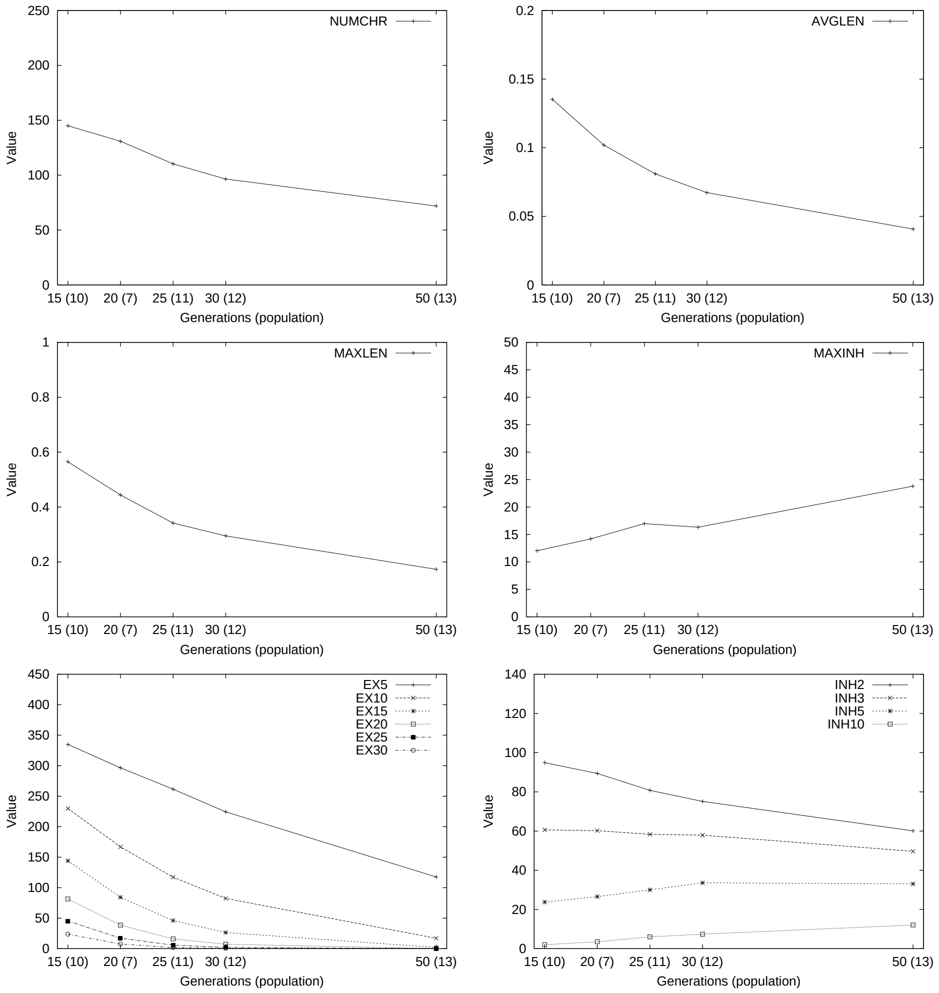

Figure 6.5: The effect of the number of generations since foundation on sharing statistics measuring the extent of conserved chromosomal segments in chromosome class 1 (pseudocontrol chromosomes of random individuals). The populations (number of generations and population number) are indicated on the horizontal axis. The vertical axis show the values of the statistics NUMCHR and AVGLEN (at the top), MAXLEN and MAXINH (in the middle), and EXn and INHn (at the bottom).

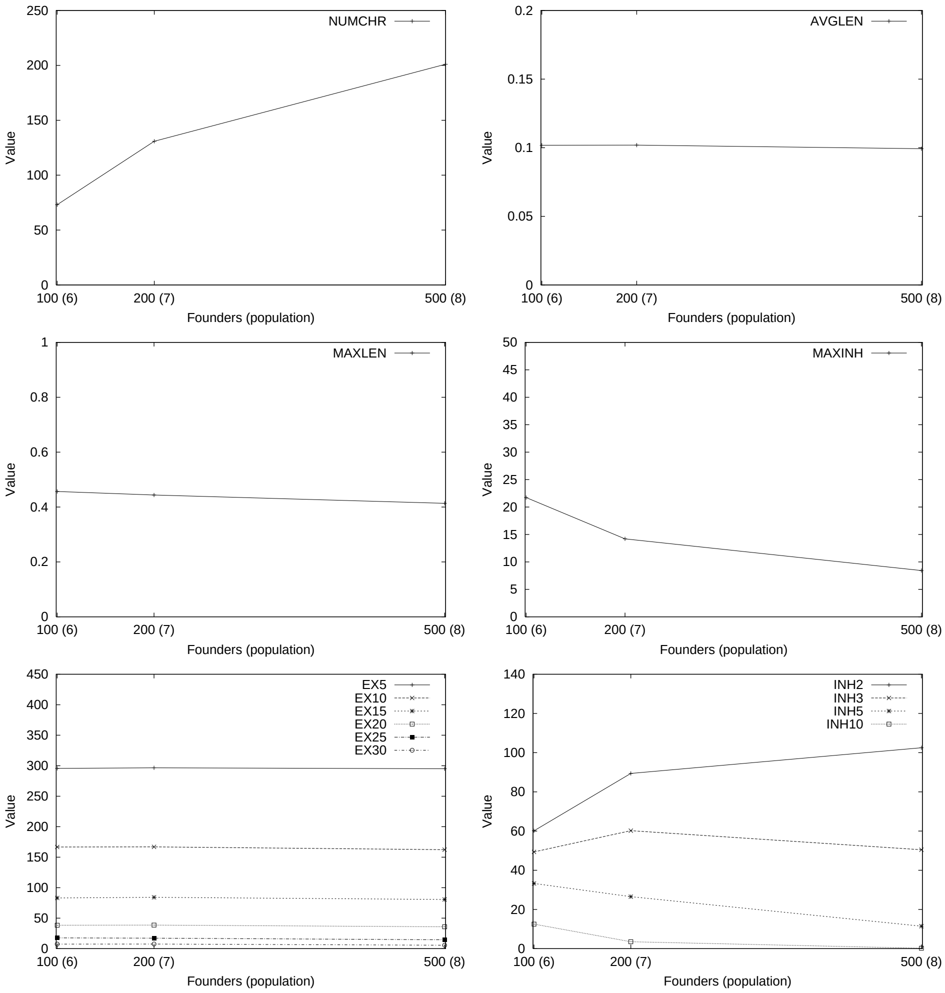

Figure 6.6: The effect of the number of initial founders on sharing statistics measuring the extent of conserved chromosomal segments in chromosome class 1 (pseudocontrol chromosomes of random individuals). The populations (number of founders and populations number) are indicated on the horizontal axis. The vertical axis show the values of the statistics NUMCHR and AVGLEN (at the top), MAXLEN and MAXINH (in the middle), and EXn and INHn (at the bottom).

Chapter 7

Detection of disease–marker association

One of the goals of this work is to help in developing techniques and measures for gene mapping. In this chapter, we use the simulator software to evaluate different measures of linkage disequilibrium. The choice of such a method is an important part in the overall methodology. First, we review some linkage disequilibrium measures and localization methods based on linkage disequilibrium (Sections 7.1 and 7.2). Then we present a collection of test populations, marker maps, and disease models and use them in an empirical study on the performance of LD measures and methods in the first stage of a genome scan (Sections 7.4 and 7.5).

7.1 Biallelic measures

Various measures for linkage disequilibrium have been proposed in the literature. The performance of these statistics has been analyzed in studies such as [Lew88], [DeR95], and [Guo97], and summarized in [Jor00]. Devlin and Risch [DeR95] examined popular linkage disequilibrium measures using simulation methods, whereas Guo [Guo97] derived analytical expressions for the expected performance. Some of the commonly used measures, compared in [Guo97] can be expressed as functions of the disease allele frequency p_d, the frequency of the other (normal) allele p_n, the conditional probabilities $p_{1|d}$ and $p_{2|d}$ for carrying allele 1 or 2 in the marker locus given that an individual has a disease allele in the disease locus of the same chromosome, and the respective probabilities $p_{1|n}$, or $p_{2|n}$ for carrying allele 1 or 2 on the condition that there is a normal allele in the disease locus of the same chromosome. These measures are summarized below:

$$
\begin{aligned}
&D = p_d p_n (p_{1|d} p_{2|n} - p_{2|d} p_{1|n});\\
&D' = \tfrac{D}{D_{max}},\ \text{where } D_{max} = \begin{cases} \min(p_d p_2, p_n p_1) \text{ when } D > 0 \\ \min(p_d p_1, p_n p_2) \text{ when } D < 0 \end{cases};\\
&\Delta = \tfrac{D}{\sqrt{p_1 p_2 p_d p_n}};\\
&\delta = \tfrac{D}{p_d p_n p_{2|n}};\\
&d = \tfrac{D}{p_d p_n};\\
&Q = \tfrac{D}{p_d p_n (p_{1|d} p_{2|n} + p_{2|d} p_{1|n})}.
\end{aligned}
$$

These six biallelic measures are included in the example analysis of this chapter. The measures consider biallelic marker loci, and require lumping together all but one of the alleles and thus omitting some information contained in the observed allele frequencies. For instance, consider a marker with 4 alleles with frequencies of 0.4, 0.3, 0.2, and 0.1. The marker could then be made biallelic by, for example, choosing the first allele (with frequency 0.4) as allele 1 and lump together the three remaining alleles as allele 2 (with frequency 0.6).

The computation of the values of these measures is straightforward, and no computational problems are associated to it.

7.2 Multiallelic measures

Since we want to focus on contingency table based measures, we here consider the commonly used multiallelic χ^2 measure, and also test the applicability of the likelihood ratio chi-squared statistic G^2, and the mutual information statistic M for the purpose of finding prominent markers in the first stage of a genome scan.

All three measures can be calculated from an $m \times n$ contingency table, where m denotes the number of different alleles in the first and n in the second marker locus, respectively. The selected multiallelic measures are written as

$$
\begin{aligned}
\chi^2 &= N \textstyle\sum_{i=1}^{m} \sum_{j=1}^{n} \frac{(p_{i|j} p_j - p_i p_j)^2}{p_i p_j};\\
G^2 &= 2N \textstyle\sum_{i=1}^{m} \sum_{j=1}^{n} p_{i|j} p_j \ln[\frac{p_{i|j} p_j}{p_i p_j}];\\
M &= \textstyle\sum_{i=1}^{m} \sum_{j=1}^{n} p_{i|j} p_j \ln[\frac{p_{i|j} p_j}{p_i p_j}],
\end{aligned}
$$

where p_i denotes the frequency of allele i in the first locus, p_j the frequency of allele j in the second locus, and $p_{i|j}$ is the conditional probability that an individual has allele i in the first locus given that he/she has allele j

in the second locus of the same chromosome. All measures are calculated from an $m \times n$ contingency table, where m denotes the number of different alleles in the first and n in the second marker locus, respectively. When a marker is compared against a disease status, the range of j in the above-mentioned formulas can be replaced by $j \in \{d, n\}$, where d corresponds to the disease allele and n to the normal allele. N denotes the total number of chromosomes.

The computation of the contingency table based multiallelic measures is fast, as was the case with biallelic measures. One of the advantages of using the biallelic measures introduced in Section 7.1 is that they generally do not suffer from problems caused by small allele frequencies. With multiallelic measures the situation may be different. For example, the multiallelic χ^2 test statistic calculated from an $m \times n$ contingency table is asymptotically χ^2 distributed with $(m-1)(n-1)$ degrees of freedom under complete linkage equilibrium. However, it may not be sufficient to rely on this asymptotic distribution, if there are small frequencies in the contingency table, as can be the case when the sample size is very limited.

The single-locus measures can easily be expanded to haplotypes of adjacent markers. As a consequence, the number of cells in the corresponding contingency table increases and the problem of small frequencies may become more serious. To overcome the problem, the distribution of the statistics under null hypothesis may have to be created from scratch each time using simulation methods.

7.3 Populations and cost models

For the purpose of analyzing the effect of different historical parameters, a collection of population descriptions, and cost models is introduced. The idea of this collection is to provide a variety of realistic cases to allow comparison of different measures and methods.

Five sets of population history descriptions are introduced here, labeled as BASIC, IMM, DEATH, STABLE, and LARGE. The parameters used are summarized in Figure 7.1.

Parameter	Year range	Age	BASIC	IMM	DEATH	STABLE	LARGE
Starting year			1500	1500	1500	1000	1000
Ending year			2000	2000	2000	2000	2000
Population size	1000—1499		n/a	n/a	n/a	600 ⇒ 1,000	10,000 ⇒
	1500—1505		300 ⇒ 320	300 ⇒ 320	300 ⇒ 320	1,000 ⇒	:
	1505—1510		320 ⇒ 350	320 ⇒ 350	320 ⇒ 350	:	:
	1510—1756		350 ⇒	350 ⇒	350 ⇒ 3,300	:	:
	1757—1763		:	:	3,300 ⇒ 1,400	:	:
	1764—1780		:	:	1,400 ⇒ 1,200	:	:
	1781—1900		⇒ 12,000	⇒ 13,200	1,200 ⇒ 12,000	⇒ 15,000	:
	1900—1950		12,000 ⇒ 30,000	13,200 ⇒ 33,000	12,000 ⇒ 30,000	15,000 ⇒ 45,000	:
	1950—2000		30,000 ⇒ 100,000	33,000 ⇒ 110,000	30,000 ⇒ 100,000	45,000 ⇒ 150,000	⇒ 1,000,000
Age structure	Initial population	0	0.03	0.03	0.03	0.03	0.03
		1	0.12	0.12	0.12	0.12	0.12
		2—5	0.15	0.15	0.15	0.15	0.15
		6—15	0.24	0.24	0.24	0.24	0.24
		16—45	0.40	0.40	0.40	0.40	0.40
		46—75	0.06	0.06	0.06	0.06	0.06
	Immigrants	0	0.03	0.03	0.03	0.03	:
		1	0.22	0.22	0.22	0.22	:
		2—5	0.21	0.21	0.21	0.21	n/a
		6—15	0.24	0.24	0.24	0.24	:
		16—30	0.20	0.20	0.20	0.20	:
		31—45	0.10	0.10	0.10	0.10	:
Death risk	until 1756	0—4	:	:	0.0689	:	:
		5—14	:	:	0.0083	:	:
		15—34	:	:	0.0081	:	:
		35—64	:	:	0.0216	:	:
		65—79	:	:	0.0883	:	:
		79—100	:	:	0.0883 → 1	:	:
	1757—1763	0—4	0.0689	0.0689	0.1000	0.0689	0.0689
		5—14	0.0083	0.0083	0.0200	0.0083	0.0083
		15—34	0.0081	0.0081	0.0100	0.0081	0.0081
		35—64	0.0216	0.0216	0.0300	0.0216	0.0216
		65—79	0.0883	0.0883	0.1200	0.0883	0.0883
		79—100	0.0883 → 1	0.0883 → 1	0.1200 → 1	0.0883 → 1	0.0883 → 1
	1764—1918	0—4	:	:	0.0689	:	:
		5—14	:	:	0.0083	:	:
		15—34	:	:	0.0081	:	:
		35—64	:	:	0.0216	:	:
		65—79	:	:	0.0883	:	:
		79—100	:	:	0.0883 → 1	:	:

Parameter	Year range	Age	BASIC	IMM	DEATH	STABLE	LARGE
Death risk (continued)	1919—1927	0—4	0.0505	0.0505	0.0505	0.0505	0.0505
		5—14	0.0057	0.0057	0.0057	0.0057	0.0057
		15—34	0.0067	0.0067	0.0067	0.0067	0.0067
		35—64	0.0180	0.0180	0.0180	0.0180	0.0180
		65—79	0.0780	0.0780	0.0780	0.0780	0.0780
		79—100	0.0780 → 1	0.0780 → 1	0.0780 → 1	0.0780 → 1	0.0780 → 1
	1928—1937	0—4	0.0320	0.0320	0.0320	0.0320	0.0320
		5—14	0.0030	0.0030	0.0030	0.0030	0.0030
		15—34	0.0053	0.0053	0.0053	0.0053	0.0053
		35—64	0.0143	0.0143	0.0143	0.0143	0.0143
		65—79	0.0676	0.0676	0.0676	0.0676	0.0676
		79—100	0.0676 → 1	0.0676 → 1	0.0676 → 1	0.0676 → 1	0.0676 → 1
	1938—1945	0—4	0.0180	0.0180	0.0180	0.0180	0.0180
		5—14	0.0019	0.0019	0.0019	0.0019	0.0019
		15—34	0.0034	0.0034	0.0034	0.0034	0.0034
		35—64	0.0126	0.0126	0.0126	0.0126	0.0126
		65—79	0.0659	0.0659	0.0659	0.0659	0.0659
		79—100	0.0659 → 1	0.0659 → 1	0.0659 → 1	0.0659 → 1	→ 1
	1946—1968	0—4	0.0040	0.0040	0.0040	0.0040	0.0040
		5—14	0.0007	0.0007	0.0007	0.0007	0.0007
		15—34	0.0015	0.0015	0.0015	0.0015	0.0015
		35—64	0.0109	0.0109	0.0109	0.0109	0.0109
		65—79	0.0641	0.0641	0.0641	0.0641	0.0641
		79—100	0.0641 → 1	0.0641 → 1	0.0641 → 1	0.0641 → 1	0.0641 → 1
	1969—2000	0—4	0.0012	0.0012	0.0012	0.0012	0.0012
		5—14	0.0002	0.0002	0.0002	0.0002	0.0002
		15—34	0.0010	0.0010	0.0010	0.0010	0.0010
		35—64	0.0074	0.0074	0.0074	0.0074	0.0074
		65—79	0.0452	0.0452	0.0452	0.0452	0.0452
		79—100	0.0452 → 1	0.0452 → 1	0.0452 → 1	0.0452 → 1	0.0452 → 1
Marriage age		18—32	0.9	0.9	0.9	0.9	0.9
Max. age of pregnancy			44	44	44	44	44
Immigration rate	until 1800		0.01	0.03	0.01	0.01	0
	1801—2000		:	0.03 → 0.06	:	:	:

Figure 7.1: Population specifications used as a basis of pedigree generation. A single arrow (→) indicates that the values are interpolated linearly, and a double arrow (⇒) that they are interpolated exponentially.

Population BASIC is a typical genetic isolate with a small initial population and rapid expansion after the initial bottleneck. This population description has much resemblance with the Finnish Kainuu population that was founded in the middle of the 16th century [KMN94], and it has expanded to its current population size of about 90,000 inhabitants (end of year 2000) [Sta02]. Another population specification, IMM, is a variant of this population. It has a high rate of incoming immigrants, which is compensated by lower birth rates. Population DEATH represent a population that has suffered from hard times, resulting in increased mortality in the 18th century. The increased immigration causes a mild drop in the population history, which creates an additional bottleneck. The boost in birth rates after the difficult years, however, ensures that the population does not become extinct but, instead, continues in a path of rapid expansion. The last of the four populations, named STABLE, differs from the first three populations in the respect that the initial foundation has occurred as early as 1,000 years ago. The population has then remained stable for a period of many hundred years until a rapid expansion has taken place in the past few centuries. The fifth population, LARGE, is also 1,000 years old, but both the initial and final population sizes have been much larger than in the other four population histories. Figure 7.2 shows how the size of each population has varied in course of time.

We focus on analyzing the cost of the initial stage of a genome scan. Three cost models will be used in the following experiments. The first model, C_1, will be based on the following three coarse assumptions:

1. Provided that the false-positive rate of the initial stage equals 0.1, it is acceptable to leave one half of the intended total budget to the subsequent stages.
2. The fixed costs of the remaining stages equal 30 per cent of the cost of the initial stage.
3. Assuming an overall power of the study of 0.9, it would be acceptable to invest an extra 30 per cent to obtain full power.

Denote the cost of the stage 0 by N, where N corresponds to $\pi_0^f + \pi_0^l$. Now, the fixed cost constant, π_1^f, is equal to $0.3N$, and the length-specific cost π_1^l is equal to $0.7N/0.1 = 7N$. The cost of a miss is estimated from equation

$$2N + (1 - 0.9)\pi^m = 2N + (0.3 \times 2N),$$

resulting in $\pi^m = 6N$. Thus, the overall cost function can be written as

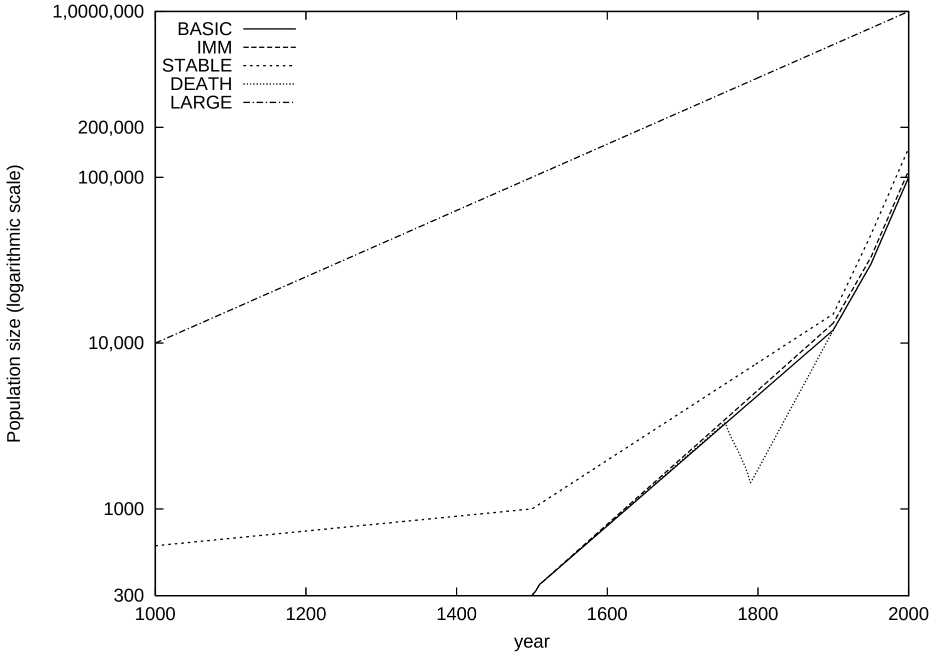

Figure 7.2: Population size as a function of time (horizontal axis) in the five population histories (BASIC, IMM, DEATH, STABLE). A logarithmic scale is used in the vertical axis, which corresponds to population size.

$$C_1 = N + 0.3N + 7NP_1 + 6N(1 - F(P_1)) = 7.3N + 7NP_1 - 6NF(P_1).$$

Cost model C_1 is based on considerations about the 'realistic' expected costs of the genome scan. Two additional cost models, C_2 and C_3 are provided. The former is based on the assumption that the relative cost of false positive regions, i.e. the length-specific cost of the regions surviving from the initial stage, is higher and the cost of miss lower, while the cost caused by a possible miss is emphasized in the latter model. All three cost models are summarized in Table 7.1. For simplicity, constant N is from now on replaced by unity.

$$\begin{aligned} C_1 &= 7.3 + 7P_1 - 6F(P_1); \\ C_2 &= 6.3 + 17P_1 - 5F(P_1); \\ C_3 &= 10.3 + 7P_1 - 9F(P_1). \end{aligned}$$

Table 7.1: The three cost models (C_1, C_2, and C_3) used in the experiments. The first cost model C_1 corresponds to an educated guess for a moderate-size genotyping facility, and the latter two models, C_2 and C_3, are modifications of it, the former emphasizing the cost of false positive signals, and the latter penalizing heavily for missing the true disease locus. The false positive rate corresponding to the transition from stage 0 (whole genome) to stage 1 (the genomic regions qualifying from the initial analysis stage) is denoted by P_1, and the corresponding power is denoted by $F(P_1)$.

7.4 Single source, fixed prevalence

7.4.1 The simulation procedure

We first generate populations according to the first four population descriptions of Section 7.3 (BASIC, IMM, DEATH, and STABLE). Then we simulate inheritance of chromosomal segments, diagnosing, sampling, and assigning a marker map. As a consequence, we get artificial haplotype data sets related to various population histories, sample sizes, and disease models. The disease models include a 'null model' with no relationship between the genes and a disease, which allows the computation of false-positive rates as defined in Section 5.2.1. The performance is then compared by applying the cost models C_1, C_2, and C_3 specified in Section 7.3. Figure 7.3 shows the pseudocode representation of the simulation procedure.

7.4.2 Disease models, marker maps, and sample sizes

The simulation procedure is used to model a single disease-contributing locus at a time, but the etiologic heterogeneity of the disease can be modeled by decreased phenocopy rate. The proportion of individuals carrying the mutation are set between 5 and 20 per cent. As a consequence, the corresponding first-order relative risks (risk of a first-order relative to be affected divided by the population prevalence of the disease [SRR78][Ris90b]) are between 1.2 and 4.1.

In addition to the 'true' disease models, we create data sets under a 'null' model containing healthy individuals only. These data sets serve as a

```
for each history h ∈ {BASIC, IMM, DEATH, STABLE} do begin
    for i:=1 to N do begin
        generate pedigree;
        generate segment data;
        for each disease model d do begin
            for each sample size s do begin
                collect sample;
                for each map m do begin
                    for each measure t do begin
                        compute observed value of statistic;
                        compute related false positive rate;
                    end;
                end;
            end;
        end;
    end;
    compute power function F_t(P_1|d, s, m, h);
    for each cost model c do begin
        compute minimum cost C_t(d, s, m, h, c);
    end;
end;
```

Figure 7.3: A pseudocode representation of the simulation procedure for creating artificial data sets, measuring the localization performance and computing the minimum cost of a setup. The computation of the power function is described in detail in Section 7.4.3.

negative control. The samples consist of 100 and 200 affected individuals, accompanied by a similar amount of pseudocontrol chromosomes, which are comprised from nontransmitted parental segments of the affected offspring. Successful sampling of disease carriers requires a sufficient amount of disease carriers in the population, so the genetic prevalence will in this context be fixed to a moderate value of 0.015.

In the experiments of this chapter, instead of fixing the allele frequencies, we fix the *information content* of each marker. The information content of a marker with n alleles is defined as a function of allele frequencies p_i

$(i \in \{1, \dots, n\})$:

$$PIC = 1 - \sum_{i=1}^{n} p_i^2 - \sum_{i=1}^{n} \sum_{j=i+1}^{n} 2p_i^2 p_j^2.$$

The information content of a marker is equal to the probability that the genotype of an offspring allows deduction of which allele an individual has inherited from his/her parent [Ott99].

As can be seen from the definition of PIC above, the information content of a marker is high, if it has several alleles with equal allele frequencies. The theoretical upper limit for the PIC of an infinitely informative marker is 1. Ott and Rabinowitz [OtR97] have demonstrated the resulting localization power to be better for marker maps with higher information content.

The information content of an entire marker map, consequently, depends on two factors: marker spacing and the information content of individual markers. Likewise, the resulting localization power should be better for marker maps with higher information content.

In the experiments, we fixed the PIC values of the microsatellite markers to 0.7. The map of uniformly spaced microsatellite markers with desired PIC were created by using a brute-force algorithm. The idea is to create large numbers of allele frequency sets using a heuristic algorithm and select the first that meet the criterion of the PIC deviating no more than a maximum allowed offset (here 0.01) from the targeted value. This process is then repeated until the desired number of markers are constructed. As a consequence, the resulting marker map consists of equidistantly located markers of similar quality, and no selection bias of any kind is introduced in the selection of the markers.

For a marker, we proceed as follows. Founder allele frequencies within each marker candidate are created by randomly selecting real numbers x and y between 0 and 1, and, if the resulting pair (x, y) satisfies inequality $y < (x-1)^4$, the value of x is approved as a frequency; otherwise x and y are rechosen. Allele frequencies are computed until their sum exceeds 1, after which the frequency of the last allele is cut down by as much as the sum departs from unity. This is just one of the numerous possibilities to create uniformly spaced markers that have constant PIC but have the property that their allele numbers and frequencies vary and have a scent of 'real' microsatellite markers as observed in genome mapping projects. The exact shape of the constraining inequality for allele frequencies has no direct interpretation, and from various alternatives, it was chosen because of the desired qualities (subject to human evaluation) of the resulting allele frequencies. Since the main focus is on the use of single nucleotide polymorphisms, a single, yet very typical, PIC value of 0.7 is used in the experiments with microsatellites.

Single nucleotide polymorphisms that utilize chip-based microarrays are soon commercially available [ZAH98]. In the experiments with SNP maps, the frequency of the more common allele is specified to values of 0.5 and 0.75 (corresponding PIC values 0.305 and 0.375), and the SNP spacing is set to 1 and 2 cM. Because of the emergence of dense SNP maps, the question about the relative performance of microsatellite and SNP markers is an interesting one, and the subject has been analyzed for single-gene diseases in isolated populations in [XiJ99]. It was observed, however, that the population frequencies of the associated alleles have a superior effect on the power and accuracy of gene localization (by TDT and linkage disequilibrium mapping) compared to the effect of marker type [XiJ99].

7.4.3 Computation of the power function

The experiments were conducted using the four population models, BASIC, IMM, DEATH, and STABLE. For each population, 100 iterations were performed. In each iteration, 2 different sample sizes, 3 disease models and 7 maps (corresponding to unique combinations of PIC and spacing) were used, each producing a different artificial data set. For each set the value of each measure was computed for all observed haplotypes of 1, 2, 3, and 4 adjacent markers. The power vs. false-positive function for a given setup was computed using the following algorithm:

1. Repeat steps 2–4 for each iteration:

2. Find marker M_i that is located immediately on the left of the true disease locus d. Variable i denotes the marker position.

3. Find the maximum value T of the test statistics computed for haplotypes that overlap or are immediately adjacent to the disease locus d, i.e. $T := max\{t(M_{i-h+1}, h), \ldots, t(M_{i+1}, h)\}$, where $t(M, h)$ denotes the (maximum) value of a test statistic computed from a haplotype of length h, starting from marker M.

4. Compare value T to the reference distribution that is obtained by merging the data sets created under null hypothesis from all iterations and computing the value of T for all intervals as indicated in the previous step. Denote the proportion of intervals where value T exceeds the values in the reference distribution by p^k. This is the false-positive rate for the current iteration, provided that the observed value of test

statistic is used as the critical level for accepting an interval to the next stage.

5. As a result of K repetitions, we have K false-positive rates. For each of these false-positive rates p^k obtained, the corresponding power $F(p^k)$ is the proportion of iterations j, where $p^j < p^k$, i.e. the proportion of cases where the false-positive rate p^k was sufficiently high to allow the true disease-containing interval to be selected.

7.4.4 Results

The power function for different biallelic measures are shown in figure 7.4, when allelic association was measured. It can be seen that the differences between the four measures (χ^2, Δ, D, δ) are small, whereas the Q and D' are not suitable for localization purposes. There are no systematic differences between the populations.

Figure 7.5 displays how the power vs. false positive curve changes as the length of haplotype changes from 1 to 4 consecutive markers, when Δ statistic has been used as a measure of association. Increasing the length of haplotype results in increased power until the length of 3 markers (6 cM) is reached, after which the power starts to decrease. This may indicate that beyond this length, the frequencies of conserved haplotypes become too small for optimal localization.

For each setup, the most cost-effective measure is displayed in tables 7.2, 7.3, 7.4, and 7.5. In the tables, S indicates an SNP and M a microsatellite marker. The best measures seem to be Δ, δ, and D, but the optimal choice varies to some extent between setups, which may be a consequence of the fact that the differences between the performance of the measures are small.

Next, we check how the optimal power/false positive ratio changes as a function of sample size and marker map design. Figure 7.6 shows optimal values for statistic Δ, when cost model C_1 has been selected. Based on the results from all four populations, we draw the conclusion that the proportion of mutation carriers in the affected sample has an effect superior to the effects of marker map design and total sample size. In all cases, the two sample schemes with the highest carrier contents (60-140 and 30-70) produced good results (power always at least 0.60 and false-positive rate always less than 0.15, irrespective of the map). On the other hand, the localization performance was appalling with the lowest carrier content and smaller sample size (10-90), but significantly better when the sample size was doubled. The differences between the four populations were rather small, possibly indicating that after the standardization of the sample com-

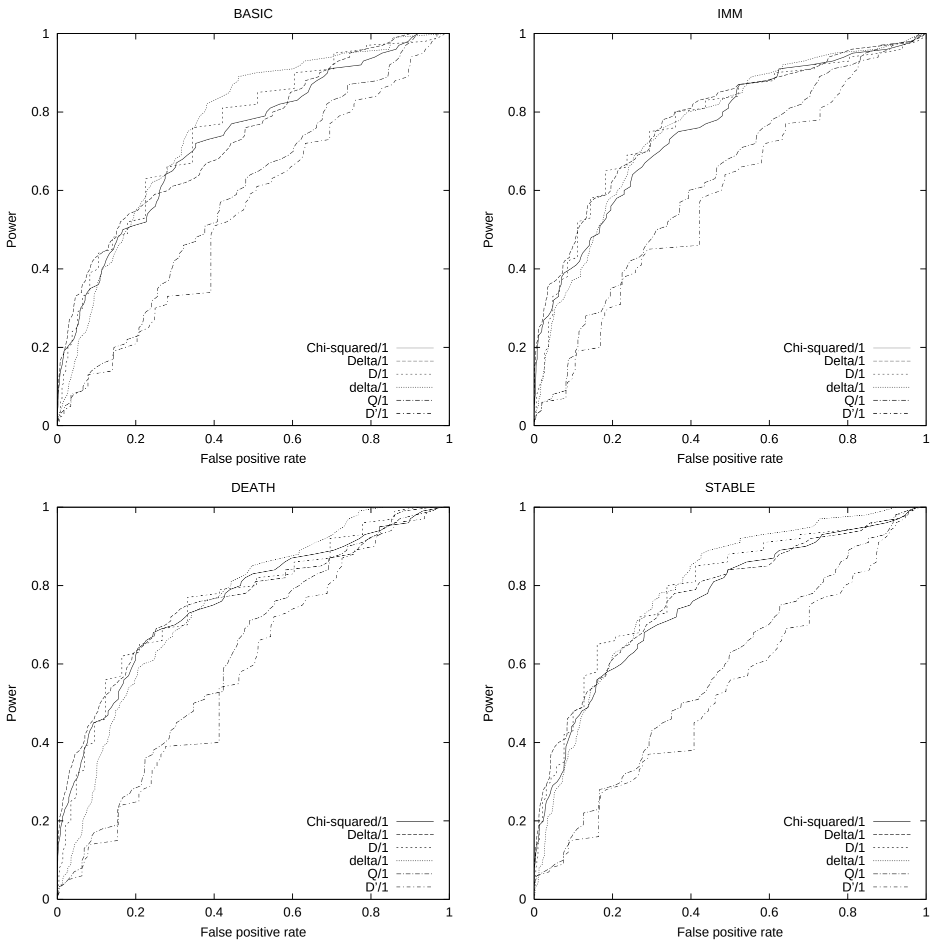

Figure 7.4: Power vs. false positive rates for allelic association (haplotypes of length 1) using different biallelic measures (marker spacing 2 cM, information content 0.7, sample of 20 mutation-carriers and 80 phenocopies).

position (number of affected carriers, number of phenocopies, and number of controls) the events in the population history have only a small effect on the localization power.

7.5 Multiple sources, variable prevalence

In the experiments of Section 7.4 the prevalence of the disease, caused by a mutation inherited from a common ancestor, had been fixed to a value of

BASIC																
Sample		Map			Cost model 1				Cost model 2				Cost model 3			
#a	#c	Type	Info.	Sp.	Meas.	F	fp	C	Meas.	F	fp	C	Meas.	F	fp	C
10	90	MS	0.7	2	$\Delta/3$	0.66	0.41	6.18	$\Delta/4$	0.03	0.00	6.15	$\Delta/3$	0.66	0.41	7.20
10	90	MS	0.7	3	Q/4	0.50	0.32	6.57	Q/1	0.02	0.00	6.21	$\chi^2_n/2$	0.93	0.76	7.26
10	90	MS	0.7	5	$\chi^2_2/2$	0.39	0.24	6.64	$\chi^2_n/1$	0.09	0.01	6.11	$\delta/4$	0.83	0.66	7.47
10	90	SNP	0.5	1	$\delta/3$	0.70	0.42	6.03	Mi/2	0.04	0.00	6.18	$\delta/4$	0.93	0.69	6.77
10	90	SNP	0.5	2	$\Delta/2$	0.48	0.32	6.65	$\delta/3$	0.05	0.01	6.19	$\delta/3$	0.77	0.63	7.78
10	90	SNP	0.75	1	$\Delta/2$	0.59	0.39	6.46	$\chi^2_n/2$	0.03	0.00	6.20	$\Delta/2$	0.85	0.65	7.19
10	90	SNP	0.75	2	Mi/4	0.71	0.51	6.64	Mi/4	0.01	0.00	6.25	mi/4	0.71	0.51	7.51
20	80	MS	0.7	2	$\Delta/4$	0.93	0.11	2.49	$\delta/4$	0.80	0.02	2.70	$\Delta/4$	0.93	0.11	2.70
20	80	MS	0.7	3	$\delta/4$	0.69	0.06	3.58	$\Delta/4$	0.60	0.03	3.82	$\Delta/4$	0.81	0.20	4.40
20	80	MS	0.7	5	D/2	0.78	0.26	4.42	$\Delta/2$	0.47	0.05	4.87	$\Delta/2$	0.88	0.39	5.08
20	80	SNP	0.5	1	$\delta/4$	0.90	0.18	3.16	$\delta/4$	0.67	0.04	3.58	$\delta/4$	0.90	0.18	3.46
20	80	SNP	0.5	2	$\delta/4$	0.72	0.14	3.93	$\delta/4$	0.44	0.03	4.56	$\delta/4$	0.79	0.21	4.65
20	80	SNP	0.75	1	D/4	0.77	0.19	4.03	$\Delta/4$	0.38	0.02	4.67	D/4	0.82	0.25	4.68
20	80	SNP	0.75	2	$\chi^2_2/4$	0.75	0.32	5.07	$\Delta/4$	0.20	0.01	5.44	D/3	0.91	0.50	5.62
30	70	MS	0.7	2	$\delta/3$	0.99	0.02	1.47	$\Delta/3$	0.97	0.00	1.48	$\delta/3$	0.99	0.02	1.50
30	70	MS	0.7	3	$\delta/4$	0.92	0.03	1.96	$\delta/4$	0.92	0.03	2.13	$\delta/4$	0.92	0.03	2.20
30	70	MS	0.7	5	D/2	0.86	0.08	2.67	$\delta/2$	0.76	0.04	3.11	$\delta/2$	0.93	0.16	3.05
30	70	SNP	0.5	1	$\delta/4$	0.97	0.02	1.62	$\delta/4$	0.97	0.02	1.80	$\delta/4$	0.97	0.02	1.71
30	70	SNP	0.5	2	$\delta/4$	0.89	0.03	2.15	$\delta/4$	0.84	0.01	2.26	$\delta/4$	0.89	0.03	2.48
30	70	SNP	0.75	1	D/4	0.94	0.11	2.43	$\Delta/4$	0.84	0.03	2.59	D/4	0.94	0.11	2.61
30	70	SNP	0.75	2	D/4	0.78	0.11	3.42	$\Delta/4$	0.65	0.04	3.66	D/4	0.78	0.11	4.08
20	180	MS	0.7	2	$\Delta/4$	0.78	0.18	3.88	$\Delta/4$	0.57	0.04	4.17	$\Delta/4$	0.78	0.18	4.54
20	180	MS	0.7	3	$\Delta/3$	0.61	0.16	4.73	$\Delta/3$	0.40	0.04	4.95	$\Delta/3$	0.74	0.28	5.59
20	180	MS	0.7	5	$\delta/4$	0.62	0.30	5.69	$\Delta/2$	0.20	0.04	5.89	$\Delta/3$	0.69	0.39	6.80
20	180	SNP	0.5	1	$\delta/4$	0.52	0.15	5.24	D/4	0.36	0.08	5.81	$\Delta/4$	0.74	0.36	6.19
20	180	SNP	0.5	2	D/4	0.46	0.19	5.88	$\Delta/4$	0.12	0.02	5.97	$\chi^2_n/4$	0.82	0.60	7.13
20	180	SNP	0.75	1	$\chi^2_2/4$	0.71	0.35	5.52	$\Delta/4$	0.17	0.02	5.87	$\chi^2_2/4$	0.71	0.35	6.39
20	180	SNP	0.75	2	$\Delta/4$	0.28	0.11	6.36	$\Delta/3$	0.13	0.03	6.11	$\chi^2_n/4$	0.97	0.85	7.51
40	160	MS	0.7	2	$\Delta/3$	0.99	0.01	1.43	$\Delta/4$	0.97	0.00	1.47	$\Delta/3$	0.99	0.01	1.46
40	160	MS	0.7	3	$\delta/3$	0.98	0.07	1.90	$\delta/4$	0.91	0.02	2.08	$\delta/3$	0.98	0.07	1.96
40	160	MS	0.7	5	$\Delta/3$	0.84	0.05	2.60	$\Delta/3$	0.84	0.05	2.93	$\Delta/3$	0.84	0.05	3.08
40	160	SNP	0.5	1	$\delta/4$	0.92	0.02	1.94	$\delta/4$	0.92	0.02	2.08	$\delta/4$	0.92	0.02	2.18
40	160	SNP	0.5	2	$\delta/4$	0.90	0.06	2.30	D/4	0.89	0.05	2.69	$\delta/4$	0.90	0.06	2.60
40	160	SNP	0.75	1	D/4	0.88	0.12	2.86	$\Delta/4$	0.71	0.02	3.17	D/4	0.88	0.12	3.22
40	160	SNP	0.75	2	D/4	0.89	0.23	3.54	$\Delta/4$	0.58	0.03	3.94	D/4	0.89	0.23	3.87
60	140	MS	0.7	2	$\chi^2_2/4$	1.00	0.00	1.30	$\chi^2_2/4$	1.00	0.00	1.30	$\chi^2_2/4$	1.00	0.00	1.30
60	140	MS	0.7	3	$\Delta/3$	1.00	0.01	1.34	D/4	0.98	0.00	1.40	$\Delta/3$	1.00	0.01	1.34
60	140	MS	0.7	5	$\Delta/3$	0.96	0.02	1.65	$\Delta/3$	0.94	0.01	1.70	$\Delta/3$	0.96	0.02	1.77
60	140	SNP	0.5	1	$\delta/4$	1.00	0.01	1.35	D/4	0.99	0.00	1.37	$\delta/4$	1.00	0.01	1.35
60	140	SNP	0.5	2	D/4	1.00	0.01	1.38	D/4	1.00	0.01	1.50	D/4	1.00	0.01	1.38
60	140	SNP	0.75	1	D/4	0.96	0.02	1.69	D/4	0.92	0.01	1.84	D/4	0.97	0.03	1.80
60	140	SNP	0.75	2	D/4	0.88	0.02	2.18	D/4	0.86	0.01	2.23	D/4	0.93	0.07	2.44

Table 7.2: The optimal measures and the corresponding power and false-positive rates for population BASIC. In the sample size column, '#a' refers to the number of affected mutation carriers, and '#c' to the number of phenocopies. In the map column, the type (Type), the information content (Info.) and the spacing (Sp.) of the markers are given. 'MS 0.7' refers to a microsatellite marker with PIC 0.7 followed by marker interval (cM). SNP refers to a SNP marker followed by the frequency of the major allele and marker interval (cM). For each cost model C_1, C_2, and C_3, the optimal measure (Meas.), optimal power (F) and false-positive rates (fp) and the corresponding minimum cost (C) are given. In the measure column, notation $\Delta/4$, for instance, indicates that the optimal measure was Δ computed from haplotypes of 4 adjacent markers.

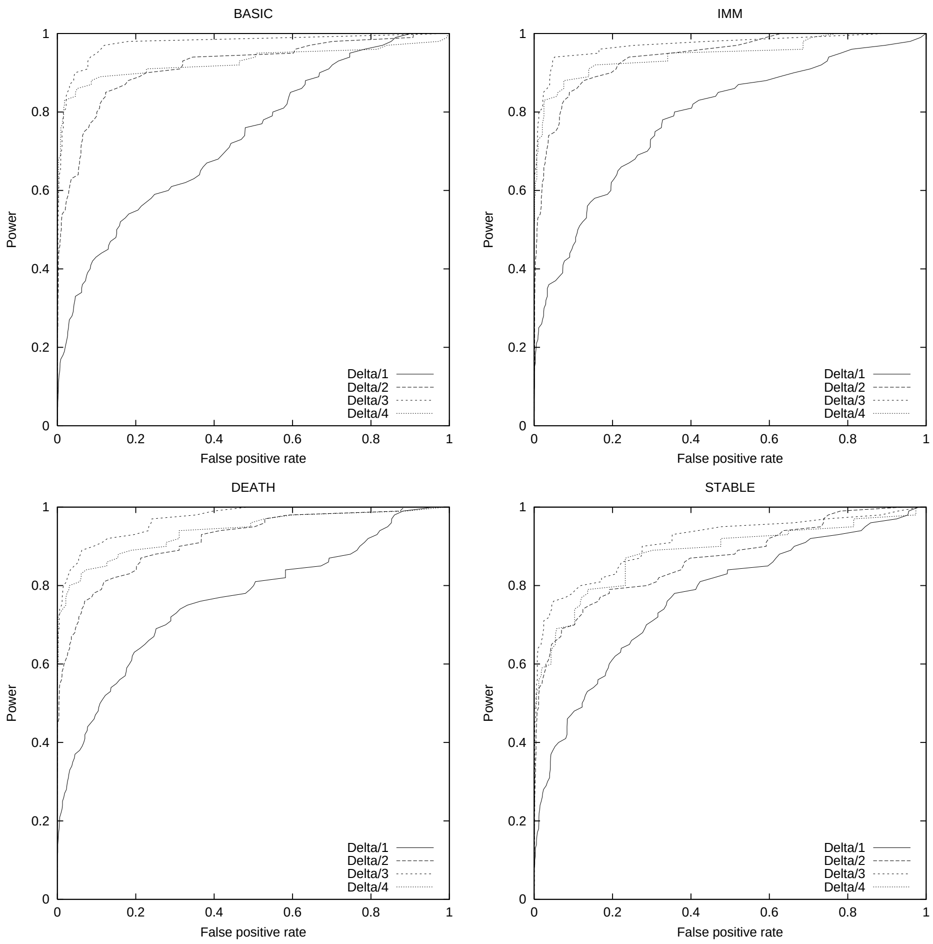

Figure 7.5: Power vs. false positive rates for haplotypes of 1, 2, 3, and 4 adjacent markers (marker spacing 2 cM, information content 0.7, sample of 20 mutation-carriers and 80 phenocopies). The Δ measure has been used as a test statistic.

0.015. Such a well-standardized setup enables efficient comparison of LD measures, but in reality the situation is often more complex. There may be several disease-predisposing loci, and each locus may contain different alleles causing the disease. Furthermore, more than one copy of the mutation may be introduced in the population, each associated to a different ancestral haplotype. In the experiments of the previous section, these complications were handled by introducing a high rate of phenocopies.

Next, we assess the localization power and cost in a setup, where the population isolate is much larger, population LARGE in Table 7.1. The

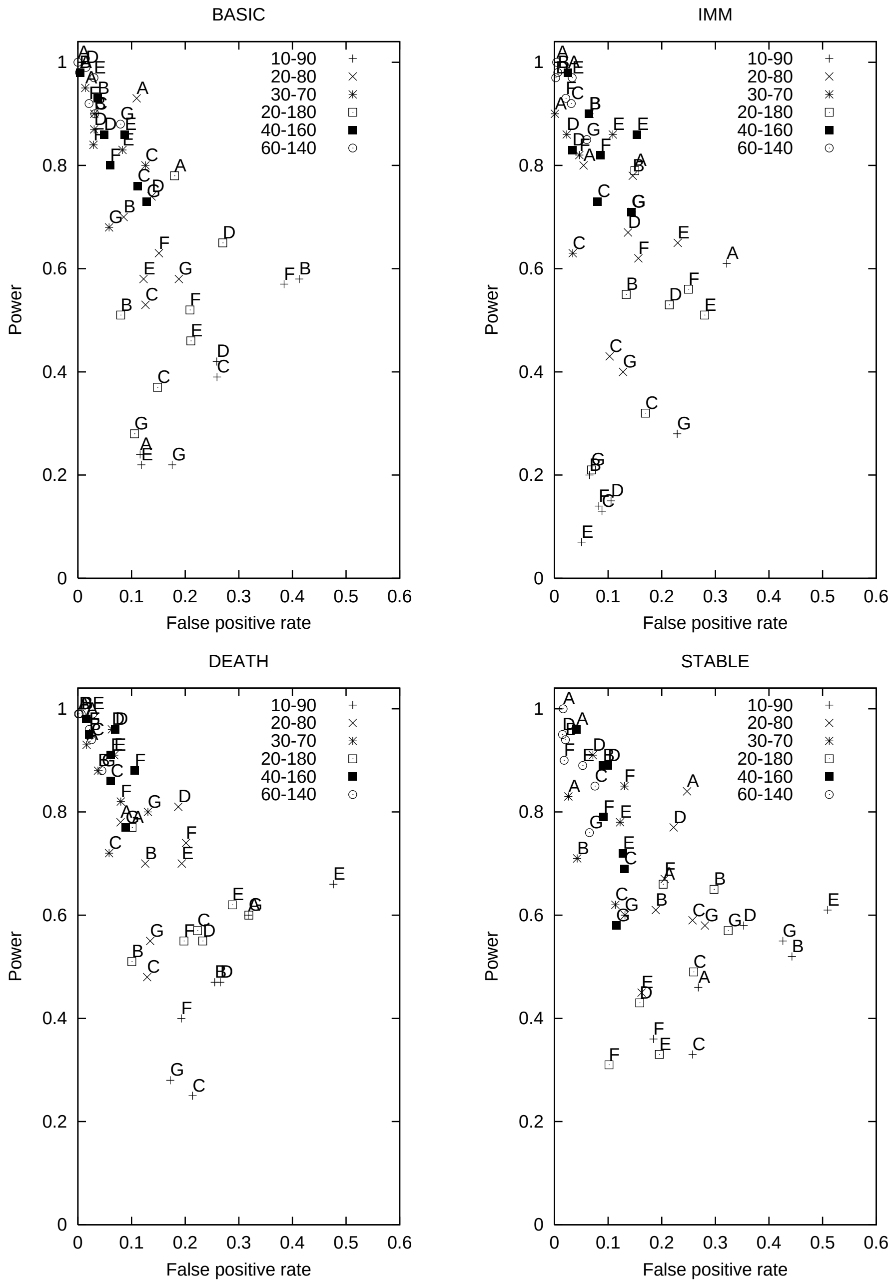

Figure 7.6: Optimal power and false positive rates for Δ measure of association using haplotypes of four adjacent markers. For a full description, see page A–1 of Appendix A.

IMM																
Sample		Map			Cost model 1				Cost model 2				Cost model 3			
#a	#c	Type	Info.	Sp.	Meas.	F	fp	C	Meas.	F	fp	C	Meas.	F	fp	C
10	90	MS	0.7	2	$\delta/4$	0.37	0.11	5.86	$\Delta/4$	0.24	0.04	5.78	$\delta/4$	0.80	0.53	6.84
10	90	MS	0.7	3	$\delta/3$	0.59	0.36	6.30	$\Delta/2$	0.07	0.01	6.06	$\Delta/3$	0.85	0.69	7.47
10	90	MS	0.7	5	$\Delta/3$	0.51	0.35	6.67	$\chi^2_n/2$	0.04	0.01	6.19	$\delta/2$	0.86	0.74	7.72
10	90	SNP	0.5	1	$\chi^2_2/4$	0.29	0.19	6.87	$D'/1$	0.02	0.00	6.25	$D'/2$	0.96	0.88	7.82
10	90	SNP	0.5	2	$\delta/4$	0.34	0.21	6.76	$D'/2$	0.03	0.00	6.22	$\Delta/3$	0.93	0.87	7.99
10	90	SNP	0.75	1	mi/1	0.65	0.48	6.73	$\delta/4$	0.04	0.00	6.18	$\chi^2_2/2$	0.92	0.80	7.64
10	90	SNP	0.75	2	$\Delta/3$	0.14	0.08	7.04	D/2	0.02	0.00	6.24	D/4	0.90	0.80	7.81
20	80	MS	0.7	2	$\delta/3$	0.89	0.06	2.41	$\delta/4$	0.75	0.02	2.82	$\delta/3$	0.89	0.06	2.74
20	80	MS	0.7	3	$\delta/3$	0.79	0.14	3.52	$\Delta/3$	0.55	0.03	4.03	$\delta/3$	0.87	0.22	4.02
20	80	MS	0.7	5	$\Delta/3$	0.63	0.20	4.92	$\Delta/3$	0.31	0.02	5.05	$\delta/3$	0.70	0.27	5.91
20	80	SNP	0.5	1	$\delta/4$	0.87	0.13	2.99	$\delta/4$	0.63	0.04	3.79	$\delta/4$	0.87	0.13	3.38
20	80	SNP	0.5	2	$\delta/4$	0.78	0.19	3.97	$\delta/4$	0.48	0.04	4.54	$\delta/4$	0.81	0.22	4.54
20	80	SNP	0.75	1	D/4	0.75	0.26	4.61	$\Delta/4$	0.45	0.05	4.88	D/3	0.85	0.36	5.15
20	80	SNP	0.75	2	D/4	0.62	0.28	5.57	$\Delta/3$	0.13	0.01	5.76	D/4	0.82	0.46	6.14
30	70	MS	0.7	2	$\delta/4$	0.96	0.03	1.72	$\delta/4$	0.93	0.01	1.77	$\delta/4$	0.96	0.03	1.84
30	70	MS	0.7	3	$\delta/4$	0.90	0.04	2.20	$\delta/4$	0.83	0.01	2.36	$\delta/4$	0.90	0.04	2.50
30	70	MS	0.7	5	$\delta/3$	0.85	0.11	2.99	$\delta/3$	0.70	0.03	3.34	$\delta/3$	0.88	0.14	3.39
30	70	SNP	0.5	1	$\delta/4$	0.95	0.02	1.72	$\delta/4$	0.93	0.00	1.71	$\delta/4$	0.95	0.02	1.87
30	70	SNP	0.5	2	$\delta/4$	0.92	0.03	1.99	$\delta/4$	0.91	0.03	2.18	$\delta/4$	0.93	0.04	2.21
30	70	SNP	0.75	1	D/4	0.92	0.09	2.39	D/4	0.85	0.05	2.91	D/4	0.92	0.09	2.63
30	70	SNP	0.75	2	D/4	0.77	0.17	3.88	D/4	0.58	0.05	4.22	$\delta/4$	0.87	0.27	4.37
20	180	MS	0.7	2	$\delta/4$	0.84	0.16	3.36	$\Delta/4$	0.60	0.03	3.73	$\delta/4$	0.84	0.16	3.84
20	180	MS	0.7	3	$\delta/4$	0.69	0.21	4.60	$\Delta/4$	0.42	0.06	5.19	$\delta/4$	0.81	0.36	5.53
20	180	MS	0.7	5	$\Delta/3$	0.55	0.23	5.61	$\Delta/4$	0.09	0.01	6.01	$\Delta/3$	0.65	0.33	6.74
20	180	SNP	0.5	1	$\delta/4$	0.46	0.12	5.37	$\delta/4$	0.34	0.05	5.45	$\delta/4$	0.75	0.38	6.24
20	180	SNP	0.5	2	$\delta/4$	0.62	0.28	5.51	$\delta/4$	0.09	0.01	5.96	$\delta/4$	0.62	0.28	6.65
20	180	SNP	0.75	1	$\Delta/4$	0.56	0.25	5.69	$\Delta/4$	0.23	0.04	5.80	D/4	0.74	0.44	6.73
20	180	SNP	0.75	2	$\Delta/4$	0.21	0.07	6.52	$\delta/4$	0.07	0.00	6.00	$\chi^2_n/4$	0.88	0.67	7.05
40	160	MS	0.7	2	$\delta/3$	0.97	0.01	1.56	$\delta/3$	0.97	0.01	1.65	D/3	0.99	0.03	1.61
40	160	MS	0.7	3	D/3	0.93	0.06	2.11	$\Delta/3$	0.87	0.02	2.34	D/3	0.93	0.06	2.32
40	160	MS	0.7	5	$\Delta/3$	0.78	0.08	3.17	$\delta/4$	0.67	0.03	3.40	$\Delta/3$	0.83	0.12	3.68
40	160	SNP	0.5	1	$\delta/4$	0.91	0.04	2.10	$\delta/4$	0.88	0.01	2.10	$\delta/4$	0.91	0.04	2.37
40	160	SNP	0.5	2	$\delta/4$	0.82	0.04	2.65	$\delta/4$	0.76	0.01	2.75	$\delta/4$	0.94	0.15	2.87
40	160	SNP	0.75	1	D/4	0.88	0.10	2.71	$\Delta/4$	0.76	0.05	3.29	D/4	0.90	0.12	3.06
40	160	SNP	0.75	2	$\Delta/4$	0.71	0.14	4.05	$\Delta/4$	0.54	0.05	4.43	D/4	0.90	0.37	4.80
60	140	MS	0.7	2	$\Delta/3$	1.00	0.00	1.31	$\Delta/3$	1.00	0.00	1.31	$\Delta/3$	1.00	0.00	1.31
60	140	MS	0.7	3	$\Delta/4$	0.98	0.01	1.46	$\Delta/4$	0.97	0.00	1.50	$\Delta/4$	0.98	0.01	1.52
60	140	MS	0.7	5	$\Delta/3$	0.91	0.01	1.91	$\Delta/3$	0.91	0.01	1.92	$\Delta/3$	0.94	0.04	2.15
60	140	SNP	0.5	1	$\delta/4$	0.98	0.00	1.43	$\delta/4$	0.98	0.00	1.42	$\delta/4$	0.98	0.00	1.49
60	140	SNP	0.5	2	$\delta/4$	0.98	0.00	1.45	$\delta/4$	0.98	0.00	1.47	$\delta/4$	0.98	0.00	1.51
60	140	SNP	0.75	1	$\Delta/4$	0.93	0.02	1.87	$\Delta/4$	0.89	0.01	1.99	$\Delta/4$	0.93	0.02	2.08
60	140	SNP	0.75	2	D/4	0.90	0.10	2.57	$\Delta/4$	0.79	0.04	2.95	D/4	0.90	0.10	2.87

Table 7.3: The optimal measures and the corresponding power and false-positive rates for population IMM. For notation, see Table 7.2.

prevalence of the disease is not fixed any more, but, instead, the frequencies of the disease-causing mutations in the initial population will be fixed. Due to random events in population and inheritance histories, some of these mutations are bound to become more common than others, while some may even become extinct. This time, the observed prevalence of the disease, as well as the unknown frequencies of the underlying mutations, varies.

7.5.1 The simulation characteristics

The creation of data sets is more complex when multiple disease sources (genetic heterogeneity) are present. This time, a large pedigree growing from 10,000 individuals up to one million individuals in a time span of one thousand years is created once, and the chromosome simulator is used to

DEATH																
Sample		Map			Cost model 1				Cost model 2				Cost model 3			
#a	#c	Type	Info	Sp.	Meas.	F	fp	C	Meas.	F	fp	C	Meas.	F	fp	C
10	90	MS	0.7	2	$\Delta/4$	0.60	0.32	5.92	Q/3	0.17	0.04	6.15	$\Delta/4$	0.78	0.52	6.95
10	90	MS	0.7	3	Q/4	0.39	0.14	5.95	$\chi^2_n/4$	0.04	0.01	6.21	$\Delta/2$	0.88	0.70	7.25
10	90	MS	0.7	5	$\chi^2_n/2$	0.54	0.35	6.49	mi/1	0.04	0.00	6.17	$\Delta/2$	0.82	0.64	7.37
10	90	SNP	0.5	1	$\Delta/4$	0.47	0.27	6.34	$\chi^2_n/3$	0.08	0.01	6.13	$\chi^2_2/4$	0.88	0.69	7.24
10	90	SNP	0.5	2	$\Delta/4$	0.66	0.48	6.67	$\Delta/4$	0.03	0.00	6.19	$\Delta/4$	0.77	0.59	7.52
10	90	SNP	0.75	1	$\Delta/4$	0.40	0.19	6.25	$\chi^2_2/4$	0.08	0.01	6.13	$\chi^2_n/4$	0.74	0.51	7.21
10	90	SNP	0.75	2	$\Delta/3$	0.44	0.26	6.46	mi/3	0.03	0.00	6.18	$\chi^2_n/3$	0.88	0.76	7.71
20	80	MS	0.7	2	$\Delta/3$	0.78	0.07	3.08	$\Delta/3$	0.71	0.04	3.35	$\Delta/4$	0.88	0.19	3.69
20	80	MS	0.7	3	$\delta/3$	0.68	0.08	3.75	$\Delta/3$	0.60	0.04	3.99	$\delta/2$	0.85	0.27	4.51
20	80	MS	0.7	5	$\Delta/3$	0.65	0.13	4.28	$\Delta/3$	0.50	0.04	4.40	$\Delta/3$	0.76	0.25	5.24
20	80	SNP	0.5	1	D/4	0.92	0.17	2.97	D/4	0.52	0.03	4.17	D/4	0.92	0.17	3.21
20	80	SNP	0.5	2	$\delta/4$	0.86	0.28	4.10	D/4	0.46	0.04	4.64	$\delta/4$	0.86	0.28	4.52
20	80	SNP	0.75	1	$\Delta/4$	0.74	0.20	4.27	$\Delta/4$	0.42	0.03	4.67	$\Delta/4$	0.74	0.20	5.05
20	80	SNP	0.75	2	$\Delta/4$	0.55	0.13	4.94	$\Delta/4$	0.42	0.06	5.29	D/4	0.86	0.48	5.92
30	70	MS	0.7	2	$\Delta/3$	0.95	0.01	1.64	$\Delta/3$	0.95	0.01	1.66	$\Delta/3$	0.95	0.01	1.79
30	70	MS	0.7	3	$\Delta/3$	0.96	0.03	1.74	$\Delta/3$	0.92	0.01	1.90	$\Delta/3$	0.98	0.05	1.83
30	70	MS	0.7	5	$\Delta/3$	0.89	0.04	2.26	$\Delta/3$	0.83	0.02	2.56	$\Delta/3$	0.89	0.04	2.59
30	70	SNP	0.5	1	$\delta/4$	0.99	0.03	1.60	D/4	0.92	0.01	1.83	$\delta/4$	0.99	0.03	1.63
30	70	SNP	0.5	2	D/4	0.88	0.02	2.14	D/4	0.88	0.02	2.20	$\delta/4$	0.91	0.05	2.46
30	70	SNP	0.75	1	D/4	0.93	0.12	2.53	D/4	0.80	0.04	2.92	D/4	0.93	0.12	2.74
30	70	SNP	0.75	2	D/4	0.86	0.15	3.22	$\Delta/4$	0.61	0.02	3.64	D/4	0.90	0.19	3.55
20	180	MS	0.7	2	$\Delta/4$	0.77	0.10	3.39	$\Delta/4$	0.63	0.06	4.10	$\Delta/4$	0.77	0.10	4.08
20	180	MS	0.7	3	$\Delta/3$	0.60	0.16	4.79	$\Delta/3$	0.44	0.05	5.02	$\Delta/3$	0.85	0.42	5.60
20	180	MS	0.7	5	$\Delta/4$	0.57	0.22	5.44	$\Delta/3$	0.15	0.00	5.61	$\Delta/4$	0.57	0.22	6.73
20	180	SNP	0.5	1	D/4	0.75	0.38	5.46	$\Delta/4$	0.13	0.01	5.89	$\delta/4$	0.84	0.48	6.08
20	180	SNP	0.5	2	$\Delta/4$	0.62	0.29	5.60	$\chi^2_n/4$	0.11	0.02	6.02	$\Delta/4$	0.62	0.29	6.74
20	180	SNP	0.75	1	$\chi^2_2/4$	0.61	0.23	5.28	$\Delta/4$	0.16	0.01	5.70	$\chi^2_2/4$	0.70	0.35	6.42
20	180	SNP	0.75	2	$\Delta/4$	0.60	0.32	5.93	$\chi^2_2/4$	0.07	0.01	6.13	$\Delta/3$	0.86	0.65	7.11
40	160	MS	0.7	2	$\Delta/4$	0.98	0.02	1.53	$\Delta/4$	0.96	0.01	1.60	$\Delta/4$	0.99	0.03	1.57
40	160	MS	0.7	3	$\Delta/4$	0.95	0.02	1.75	$\Delta/4$	0.94	0.01	1.84	$\Delta/4$	0.95	0.02	1.90
40	160	MS	0.7	5	$\Delta/3$	0.85	0.04	2.48	$\Delta/3$	0.79	0.02	2.61	$\Delta/3$	0.89	0.08	2.84
40	160	SNP	0.5	1	$\delta/4$	0.99	0.04	1.61	D/4	0.97	0.02	1.86	$\delta/4$	1.00	0.05	1.63
40	160	SNP	0.5	2	$\Delta/4$	0.91	0.06	2.27	D/4	0.85	0.03	2.56	$\Delta/4$	0.96	0.11	2.46
40	160	SNP	0.75	1	D/4	0.93	0.15	2.75	$\Delta/4$	0.72	0.03	3.15	D/4	0.93	0.15	2.96
40	160	SNP	0.75	2	$\Delta/4$	0.77	0.09	3.30	$\Delta/4$	0.61	0.02	3.66	$\Delta/4$	0.78	0.10	3.96
60	140	MS	0.7	2	D/3	1.00	0.00	1.32	D/3	1.00	0.00	1.36	D/3	1.00	0.00	1.32
60	140	MS	0.7	3	$\Delta/3$	1.00	0.00	1.31	$\Delta/3$	1.00	0.00	1.32	$\Delta/3$	1.00	0.00	1.31
60	140	MS	0.7	5	$\Delta/3$	0.99	0.01	1.40	$\Delta/3$	0.99	0.01	1.44	$\Delta/3$	0.99	0.01	1.43
60	140	SNP	0.5	1	$\delta/4$	1.00	0.00	1.31	$\delta/4$	1.00	0.00	1.33	$\delta/4$	1.00	0.00	1.31
60	140	SNP	0.5	2	$\delta/4$	0.99	0.01	1.41	$\delta/4$	0.98	0.00	1.47	$\delta/4$	0.99	0.01	1.44
60	140	SNP	0.75	1	D/4	0.99	0.03	1.54	D/4	0.97	0.01	1.66	D/4	0.99	0.03	1.57
60	140	SNP	0.75	2	D/4	0.94	0.07	2.16	$\chi^2_2/4$	0.82	0.01	2.38	D/4	0.97	0.10	2.30

Table 7.4: The optimal measures and the corresponding power and false-positive rates for population DEATH. For notation, see Table 7.2.

simulate inheritance within that pedigree.

Instead of selecting one fixed mutation source for the entire population, a different procedure is applied: a very dense map of markers is created, some of which are regular SNPs and/or microsatellites and some are biallelic markers representing potential disease-causing mutations, called 'gene-SNPs'. An individual's genotypes in selected gene-SNP loci then defines his/her risk to become affected. This probability is taken from the penetrance matrix, which is a representation of the disease model.

The data sets for the analysis are created in the following steps. In the algorithm, N_c denotes the number of chromosomes, which is two in these experiments. The algorithm can be generalized to accommodate any number of disease-predisposing loci in any number of chromosomes, but the corresponding disease models and, consequently, the localization of the

STABLE																
Sample		Map			Cost model 1				Cost model 2				Cost model 3			
#a	#c	Type	Info	Sp.	Meas.	F	fp	C	Meas.	F	fp	C	Meas.	F	fp	C
10	90	MS	0.7	2	$\Delta/4$	0.46	0.27	6.42	Q/4	0.03	0.00	6.20	$\chi_2^2/1$	0.87	0.68	7.23
10	90	MS	0.7	3	$\delta/1$	0.34	0.21	6.70	$\chi_2^2/3$	0.03	0.00	6.15	$\chi_n^2/2$	0.81	0.62	7.38
10	90	MS	0.7	5	$\chi_2^2/1$	0.23	0.14	6.91	$\chi_2^2/3$	0.04	0.00	6.13	$\chi_2^2/4$	0.91	0.81	7.78
10	90	SNP	0.5	1	$\chi_n^2/4$	0.49	0.26	6.19	$D'/1$	0.05	0.01	6.16	$\Delta/4$	0.92	0.76	7.35
10	90	SNP	0.5	2	$\chi_n^2/2$	0.23	0.10	6.61	$\chi_n^2/4$	0.03	0.00	6.17	$\chi_n^2/4$	0.76	0.62	7.82
10	90	SNP	0.75	1	$\delta/3$	0.65	0.41	6.27	Q/2	0.02	0.00	6.23	$\delta/3$	0.73	0.49	7.17
10	90	SNP	0.75	2	$\Delta/1$	0.71	0.49	6.49	Q/4	0.02	0.00	6.22	Q/1	0.74	0.52	7.29
20	80	MS	0.7	2	$\Delta/3$	0.78	0.13	3.55	$\Delta/3$	0.50	0.03	4.28	$\Delta/3$	0.79	0.14	4.18
20	80	MS	0.7	3	$\Delta/3$	0.74	0.28	4.81	$\Delta/3$	0.37	0.03	4.89	$\Delta/3$	0.86	0.43	5.57
20	80	MS	0.7	5	$\Delta/4$	0.59	0.26	5.56	$\Delta/2$	0.17	0.03	5.90	$\Delta/2$	0.71	0.37	6.48
20	80	SNP	0.5	1	$\delta/4$	0.70	0.11	3.87	$\delta/4$	0.66	0.08	4.41	$\delta/4$	0.70	0.11	4.77
20	80	SNP	0.5	2	$\delta/4$	0.60	0.24	5.37	D/4	0.19	0.03	5.89	$\delta/4$	0.60	0.24	6.57
20	80	SNP	0.75	1	$\Delta/4$	0.67	0.21	4.72	$\Delta/4$	0.26	0.01	5.23	$\Delta/4$	0.67	0.21	5.71
20	80	SNP	0.75	2	D/2	0.57	0.25	5.66	$\Delta/4$	0.21	0.01	5.50	D/2	0.57	0.25	6.95
30	70	MS	0.7	2	D/2	0.92	0.07	2.24	$\Delta/3$	0.85	0.02	2.34	D/2	0.96	0.11	2.43
30	70	MS	0.7	3	$\Delta/3$	0.93	0.11	2.51	$\Delta/3$	0.75	0.01	2.74	$\Delta/3$	0.93	0.11	2.72
30	70	MS	0.7	5	$\Delta/2$	0.81	0.09	3.10	$\Delta/2$	0.70	0.04	3.56	$\Delta/2$	0.81	0.09	3.67
30	70	SNP	0.5	1	$\delta/4$	0.91	0.05	2.17	D/4	0.82	0.02	2.48	D/4	0.93	0.07	2.38
30	70	SNP	0.5	2	$\delta/4$	0.85	0.07	2.71	D/4	0.78	0.03	2.94	$\delta/4$	0.87	0.09	3.11
30	70	SNP	0.75	1	D/4	0.80	0.08	3.06	$\Delta/4$	0.66	0.01	3.24	D/4	0.92	0.19	3.35
30	70	SNP	0.75	2	D/3	0.63	0.11	4.29	$\Delta/4$	0.44	0.03	4.55	$\delta/3$	0.87	0.38	5.15
20	180	MS	0.7	2	$\Delta/3$	0.62	0.16	4.71	$\Delta/4$	0.37	0.03	5.04	$\Delta/3$	0.68	0.22	5.69
20	180	MS	0.7	3	$\Delta/4$	0.65	0.30	5.48	$\Delta/3$	0.20	0.03	5.77	$\Delta/4$	0.65	0.30	6.53
20	180	MS	0.7	5	$\chi_2^2/2$	0.35	0.14	6.15	$\Delta/1$	0.06	0.01	6.14	$\Delta/2$	0.75	0.51	7.14
20	180	SNP	0.5	1	D/4	0.55	0.24	5.65	$\chi_2^2/4$	0.21	0.02	5.62	$D'/3$	0.80	0.52	6.75
20	180	SNP	0.5	2	mi/4	0.17	0.04	6.59	mi/4	0.17	0.04	6.19	$\chi_n^2/4$	0.78	0.61	7.58
20	180	SNP	0.75	1	$\Delta/4$	0.31	0.10	6.15	$\Delta/4$	0.20	0.03	5.89	$\Delta/4$	0.80	0.58	7.15
20	180	SNP	0.75	2	mi/3	0.59	0.31	5.93	$\Delta/3$	0.10	0.01	5.95	mi/3	0.60	0.32	7.13
40	160	MS	0.7	2	$\Delta/4$	0.96	0.04	1.83	$\Delta/3$	0.89	0.01	2.01	$\Delta/4$	0.96	0.04	1.95
40	160	MS	0.7	3	$\Delta/3$	0.90	0.07	2.37	$\Delta/3$	0.84	0.04	2.75	$\Delta/3$	0.90	0.07	2.67
40	160	MS	0.7	5	D/2	0.80	0.17	3.67	$\Delta/2$	0.57	0.03	3.99	D/3	0.89	0.26	4.14
40	160	SNP	0.5	1	$\delta/4$	0.90	0.07	2.38	D/4	0.85	0.03	2.57	$\delta/4$	0.94	0.11	2.60
40	160	SNP	0.5	2	$\delta/4$	0.87	0.21	3.53	$\delta/4$	0.62	0.05	4.05	$\delta/4$	0.87	0.21	3.92
40	160	SNP	0.75	1	D/4	0.84	0.12	3.10	$\Delta/4$	0.75	0.06	3.59	D/4	0.91	0.21	3.56
40	160	SNP	0.75	2	D/4	0.70	0.17	4.30	$\Delta/4$	0.44	0.04	4.75	D/4	0.70	0.17	5.20
60	140	MS	0.7	2	$\Delta/4$	1.00	0.02	1.41	D/3	0.98	0.01	1.50	$\Delta/4$	1.00	0.02	1.41
60	140	MS	0.7	3	D/3	0.98	0.03	1.66	$\Delta/3$	0.91	0.00	1.79	D/3	0.98	0.03	1.72
60	140	MS	0.7	5	D/2	0.90	0.05	2.26	D/2	0.88	0.04	2.51	D/2	0.96	0.11	2.43
60	140	SNP	0.5	1	$\delta/4$	0.98	0.01	1.48	$\delta/4$	0.97	0.00	1.52	$\delta/4$	0.98	0.01	1.54
60	140	SNP	0.5	2	$\delta/4$	0.96	0.05	1.92	D/4	0.87	0.01	2.09	$\delta/4$	0.96	0.05	2.04
60	140	SNP	0.75	1	D/4	0.95	0.05	1.96	$\Delta/4$	0.90	0.02	2.11	D/4	0.95	0.05	2.11
60	140	SNP	0.75	2	D/4	0.82	0.06	2.81	D/4	0.78	0.04	3.10	D/4	0.82	0.06	3.35

Table 7.5: The optimal measures and the corresponding power and false-positive rates for population STABLE. For notation, see Table 7.2.

disease genes may then become very complex.

generate pedigree;
for $i := 1$ **to** N_c **do begin**
    generate segment data;
    generate marker/SNP map;
    assign alleles and environmental covariates;
end;
for $k := 1$ **to** N **do begin**
    select disease-predisposing loci;
    diagnose;
    pick sample;
    for each measure t **do**

Genotype 1	Genotype 2	Penetrance
11	11	0.001
11	12	0.1
11	22	0.1
12	11	0.1
12	12	1.0
12	22	1.0
22	11	0.1
22	12	1.0
22	22	1.0

$$\Pi = \begin{pmatrix} 0.001 & 0.1 & 0.1 \\ 0.1 & 1 & 1 \\ 0.1 & 1 & 1 \end{pmatrix}$$

Figure 7.7: The two-chromosome, two-gene disease model in the form of a penetrance matrix (table and matrix representations). There are two disease loci, or 'gene-SNPs' which contribute to the individual's probability of becoming affected. A probability for being affected is assigned to each combination of two-locus genotypes in the model. The expected frequency of each disease allele in the founder population is set to 0.002.

        compute observed value of statistic;
        compute related false positive rate;
    end;
end;
compute power function $F_t(P)$;
for each cost model c **do begin**
    find minimum cost $C_t(c)$;
end;

7.5.2 Disease model, marker maps, and sampling strategy

A two-locus model where the two disease loci are situated in different chromosomes (Figure 7.7) is used in determining the phenotype for the members of the population in the last generation.

Two-chromosome data was created to demonstrate the capability of the simulator software to produce such data for future purposes. However, the specified model is symmetrical with respect to the two genes in the different chromosomes, so the localization power is analyzed for the first chromosome only.

For all of the analysis, a single map with 301 markers in the 100 cM area is used. This yields the maximum marker density of 0.33 cM. In addition, there are 100 randomly located gene-SNPs in each chromosome. In each iteration, two of them are randomly selected as disease-predisposing loci, and a disease-predisposing variant (disease allele) is introduced to each founder with the probability of 0.002. The penetrances corresponding to different genotypes in these two loci are listed in Table 7.7. The normal allele in the mutation locus is denoted by 1 and a mutated allele by 2. In each marker locus, the frequency of the major allele was taken from a distribution, where in 70 per cent of the cases the frequency was taken from uniform distribution between 0.5 and 0.6, and in the remaining 30 per cent from uniform distribution with endpoints at 0.6 and 0.75. The idea of this procedure is to generate a considerable number of high-quality SNPs accompanied by a few lower-quality ones.

As a consequence of the simulation procedure, the average number of affected individuals in the final population is 1786.8. This corresponds well with the expected number of affected individuals in the original population, indicating that the proportion of affected individuals has not been considerably altered by genetic drift. The average number of affected individuals who have both parents alive at the time of sampling is 834.8. Since these individuals can be used as index cases for collecting trios, this count is an upper bound for sample size in the analysis.

The sample sizes were fixed to 500, 400, 300, 200, and 100 trios, and map densities of 1/3, 2/3, and 1 cM were used. The maps were created by selecting all markers, every second, and, finally, every third marker from the original map of 301 markers. There were 2 cases, where the largest sample size of 500 trios could not be reached because of a lower number of affected individuals (489 and 439) in the final population, but these samples are included in the analysis anyhow. The cost models used in the analysis are similar to those of Section 7.4.

7.5.3 Results

The most powerful linkage disequilibrium measures as well as the corresponding power and false positive rates are shown in Table 7.6. The table indicates that a reasonable power/false positive rate can be obtained even for sample sizes of 200 or more affected individuals, if very dense SNP maps (3 per cM) are available. With larger sample sizes, fair power (0.7) can be obtained even with SNPs placed at intervals of 1 cM, which seems to indicate that use of dense SNP maps may be an attractive alternative for genome scan studies.

The optimal haplotype lengths for different measures and setups are listed in Table 7.7. The results illustrate the fact that for the most applicable measures (d, D, δ, Δ and χ^2_2) the optimal haplotype length is equal or close to the maximum value of 12 adjacent markers. The gain of using longer haplotypes seems to outweigh the possible inaccuracy in selecting prominent regions. For other measures with limited power, the optimal length greatly varies, which may be an indication of the unstable behavior

Sample		Cost model 1				Cost model 2				Cost model 3			
Size	D.	Meas.	F	fp	C	Meas.	F	fp	C	Meas.	F	fp	C
100	3	$\delta/12$	0.69	0.04	3.42	D/12	0.63	0.00	3.21	$\chi_2^2/12$	0.75	0.09	4.20
100	2	$\chi_2^2/12$	0.49	0.04	4.64	$\delta/11$	0.42	0.01	4.32	$D'/11$	0.55	0.10	6.05
100	1	$D'/12$	0.30	0.02	5.61	$D'/12$	0.30	0.02	5.06	$\chi_2^2/12$	0.32	0.03	7.65
200	3	$\delta/12$	0.81	0.02	2.61	$\Delta/12$	0.79	0.01	2.55	$\delta/12$	0.81	0.02	3.18
200	2	$\chi_2^2/12$	0.82	0.05	2.74	$\chi_2^2/12$	0.82	0.05	3.08	$\chi_2^2/12$	0.82	0.05	3.28
200	1	$\chi_2^2/12$	0.57	0.06	4.31	D/11	0.49	0.03	4.44	$\chi_2^2/11$	0.72	0.25	5.57
300	3	$\delta/12$	0.84	0.02	2.39	$\Delta/12$	0.81	0.00	2.27	$\delta/12$	0.84	0.02	2.87
300	2	$\delta/12$	0.79	0.02	2.71	$\delta/12$	0.75	0.01	2.68	$\delta/12$	0.79	0.02	3.34
300	1	$\delta/12$	0.55	0.01	4.09	$\delta/12$	0.55	0.01	3.77	$\delta/12$	0.67	0.15	5.29
400	3	D/12	0.82	0.00	2.38	D/12	0.82	0.00	2.21	$\Delta/12$	0.83	0.01	2.90
400	2	$\delta/12$	0.83	0.02	2.46	$\delta/12$	0.79	0.01	2.45	$\delta/12$	0.83	0.02	2.97
400	1	$\chi_2^2/12$	0.78	0.13	3.51	D/12	0.63	0.02	3.49	$\chi_2^2/12$	0.78	0.13	4.17
500	3	$\delta/12$	0.84	0.00	2.29	$\delta/12$	0.84	0.00	2.17	$\delta/12$	0.84	0.00	2.77
500	2	D/12	0.84	0.01	2.35	$\delta/12$	0.82	0.01	2.29	D/12	0.84	0.01	2.83
500	1	$\chi_2^2/12$	0.77	0.07	3.18	$\chi_2^2/12$	0.70	0.02	3.07	$\chi_2^2/12$	0.77	0.07	3.87

Table 7.6: The most cost-efficient linkage disequilibrium measures. Sample size corresponds to the number of affected individuals in the sample. The column labeled as 'D.' denotes marker density (markers/cM). For each cost model C_1, C_2, and C_3, the optimal measure (Meas.), optimal power (F) and false-positive rates (fp) and the corresponding minimum cost (C) are given. In the measure column, notation $D'/12$, for instance, indicates that the optimal measure for the setup was D' computed from haplotypes of 12 adjacent markers.

of these statistics.

Conclusions To conclude, the main findings of this section are that dense SNP maps can have adequate power even for the initial stage of a genome scan, and that at this scale the selection of prominent regions should always be based on haplotypes obtained from long sequences of adjacent markers. As such, the experiment acts as a proof of principle for whole-genome linkage disequilibrium mapping in an isolated population using a set of relatively dense SNP markers. Another observation is that the choice of test statistic is not crucial: any of the five contingency-table based statistics seemed to perform reasonably well.

It is interesting to compare these results to studies of Long and Langley [LoL99], where the power to detect association with multiple SNPs was analyzed for very narrow candidate regions corresponding to physical lengths of 25 to 250 kB, and consequently genetic lengths of roughly 0.0025 to 0.25 cM. In the experiments, the coalescent history of the region was modeled, including the evolution of SNP markers. There were some interesting findings in [LoL99]: first, marker-based permutation tests turned out to be more powerful than haplotype-based tests, and second, an increase in sample size had a greater effect on the power than an increase in the number of SNPs.

Cost model 1											
Sample		Measure									
Size	Density	d	D	δ	Δ	D'	G^2	MI	Q	χ_2^2	χ_n^2
100	3	12	12	12	12	12	5	5	12	12	7
100	2	12	12	12	11	11	7	7	12	12	7
100	1	12	12	12	12	12	10	10	12	12	10
200	3	12	12	12	12	12	6	6	12	12	5
200	2	12	12	12	12	12	6	6	12	12	6
200	1	11	11	11	11	12	4	4	12	12	5
300	3	12	12	12	12	2	5	5	12	12	5
300	2	12	12	12	12	12	6	6	12	12	6
300	1	12	12	12	12	12	5	5	12	12	6
400	3	12	12	12	12	3	6	6	3	12	6
400	2	12	12	12	12	4	5	5	12	12	5
400	1	12	12	12	12	3	3	3	12	12	3
500	3	12	12	12	12	4	6	6	2	12	6
500	2	12	12	12	12	3	5	5	3	12	5
500	1	12	12	12	12	1	3	3	2	12	3
Cost model 2											
Sample		Measure									
Size	Density	d	D	δ	Δ	D'	G^2	MI	Q	χ_2^2	χ_n^2
100	3	12	12	12	12	12	5	5	12	12	7
100	2	11	11	11	11	12	7	7	12	11	7
100	1	12	12	12	12	12	10	10	12	12	9
200	3	11	11	12	12	1	4	4	12	11	4
200	2	12	12	12	12	12	6	6	12	12	6
200	1	11	11	11	11	1	3	3	12	12	3
300	3	12	12	12	12	2	6	6	3	12	6
300	2	12	12	12	12	2	6	6	12	12	6
300	1	12	12	12	12	2	5	5	12	12	5
400	3	12	12	12	12	3	6	6	2	12	7
400	2	11	11	12	11	2	5	5	2	11	5
400	1	12	12	12	12	3	3	3	3	12	6
500	3	12	12	12	12	3	6	6	2	12	6
500	2	11	11	12	12	2	6	6	3	11	6
500	1	12	12	12	12	3	5	5	3	12	5
Cost model 3											
Sample		Measure									
Size	Density	d	D	δ	Δ	D'	G^2	MI	Q	χ_2^2	χ_n^2
100	1	12	12	12	12	12	8	8	12	12	7
100	2	11	11	11	11	11	12	12	11	12	10
100	3	12	12	10	12	12	10	10	12	12	10
200	1	12	12	12	12	12	5	5	12	12	6
200	2	12	12	12	12	12	5	5	12	12	6
200	3	11	11	11	11	12	4	4	12	11	5
300	1	12	12	12	12	12	5	5	12	12	5
300	2	12	12	12	11	12	6	6	12	12	5
300	3	11	11	12	12	12	6	6	12	12	6
400	1	12	12	12	12	3	5	5	3	12	5
400	2	11	11	12	12	4	5	5	12	12	5
400	3	12	12	12	12	3	3	3	12	12	3
500	1	12	12	12	12	3	6	6	3	11	6
500	2	12	12	12	12	4	5	5	12	12	5
500	3	12	12	12	12	1	3	3	12	12	7

Table 7.7: Haplotype length (1 to 12 SNP markers) that yields the lowest cost for the three cost models (C_1 at the top, C_2 in the middle, and C_3 at the bottom). For each sample size and marker density (markers/cM), the optimal haplotype length for each measure is given.

The first of these observations seems at first glance contradictory to our results. However, the different results are probably due to the different use of the term 'haplotype': whereas we refer to subhaplotypes of variable length (1, 2, 3, ... markers), Long and Langley test the association of haplotypes spanning the entire region. Even though the regions analyzed in [LoL99] are extremely narrow, the time scale is much larger, and SNP mutations (which we could ignore) have probably had a major effect in the differentiation of the marker alleles. Thus, it is not surprising that tests for haplotypes of a large number of markers may not be as powerful as those for a single marker close to the disease-predisposing variant. It might be interesting to know, if the power of the tests in [LoL99] would have increased if conducted for haplotypes of (a few) adjacent markers, as in our experiments.

The second result of [LoL99] that collecting additional samples is preferable to adding more markers is likely to be true for very densely spaced marker sets, but for the (relatively) sparse maps like ours, adding more markers is definitely of benefit comparable to increasing the sample size (see Table 7.6). The sufficient density of the marker map, however, depends on the population history, and it is evident that beyond that density, collecting new samples is indeed preferable to genotyping more markers for the existing individuals.

Chapter 8

Population history and disease gene localization

The target of this chapter is to shed light on the relationship between the population history and applicability of linkage disequilibrium mapping. In previous chapters, we started from the nature and extent (Chapter 6) of linkage disequilibrium, and moved towards analyzing different measures of linkage disequilibrium in localization of human disease genes (Chapter 7). From the measures of disease-marker association, we now move to the effect of population history, disease etiology, and study design. We utilize complex disease models and compare commonly used, advanced methods of linkage disequilibrium mapping in detecting disease genes. Specifically, the focus lies on the following questions:

1. How applicable is linkage disequilibrium mapping in the initial phase of a genome scan? Is it necessary to begin by collecting large pedigrees for linkage analysis, or is it feasible to start with nuclear families and association-based methods?

2. How do unobserved factors in population history affect the feasibility of linkage-disequilibrium mapping? These factors include hidden population substructure, population size, age, and rate of growth.

3. How does the power of the genome scan depend on the choice of mapping method, marker map density and sample collection strategy?

Due to the challenging disease models and the initial phase of a genome scan, we are in a situation, where the power of linkage disequilibrium mapping is not generally sufficient for demonstrating significant association between the disease locus and associated alleles or haplotypes. Instead, the task is to find prominent regions to focus on in the subsequent phases (see

Section 5.2.1). It will be analyzed, how well this classification between the two groups, i.e. prominent and uninteresting regions, can be done.

In the experiments of this chapter, we focus on methods that detect haplotype association to the phenotype. We assume that the true haplotypes are known, or can be constructed using the genotype information of the previous generations.

8.1 The experimental design

As gene mapping tools, we use three haplotype-based methods that utilize linkage disequilibrium between adjacent markers, namely Transmission/disequilibrium tests (TDT) [SGE93] as implemented in [KDR96], Haplotype pattern mining (HPM) [TOV00], and Haplotype relative risk (HRR) [RWC81, FaR87, TeO92] extended to multilocus haplotypes. These three methods are used as a tool of ranking different intermarker intervals in the initial stage of a genome mapping project, which is a rather unorthodox use for the methods (with a possible exclusion of HPM, which was designed with this kind of approaches in mind). We also understand that the haplotype length limit of four adjacent markers in the implementation of the TDT [KDR96] may be unfavorable for the method.

The data generation procedure was repeated for each setup, which consists of a population history description, marker map density, and disease model. The generation of chromosomal segments of affected individuals and their parents was similar to that described in Section 6.1. After creating the segment data, a map of genetic markers was introduced, and alleles were assigned to all founder individuals, and, subsequently, to the sampled individuals. For each setup, 100 data sets of 200 affected individuals and their parents are created. For each data set, this yields 400 disease-associated and 400 control chromosomes. By fixing the sample size, the effect of factors such as population history and disease model should become visible. In real experiments, decreased power would naturally be compensated by increasing the sample size.

The 100 data sets in each collection were analyzed one at a time with the three linkage disequilibrium mapping methods. For all methods, markerwise scores were computed to represent the degree of association to the phenotype. HPM directly produces marker-wise scores. For TDT and HRR we used the chi-squared value of the most strongly associated haplotype overlapping the marker. The predicted location of the disease-susceptibility gene was the marker with highest score, or the centerpoint of the plateau of markers with similar scores. Thus, for each setup, and each method, the output of the procedure was a collection of (marker, score) pairs and a point estimate for the gene location. Based on these, three statistics summarizing the performance of each method were defined:

ERR The average localization error of a method. This is computed over the 100 data sets, and defined as the average of the absolute difference between the predicted gene locus and the true gene locus. In the total absence of a genetic effect, or when using random guessing as a prediction, the expected value for this statistic is equal to one third of the region, i.e. $33,3$ cM.

FPR The false-positive rate of the method. This is the average proportion of the total region computed that needs to be searched in vain in order not to miss the true locus. It is computed by calculating the number of markers, where the score is higher than or equal to the score of the marker closest to the true disease locus and dividing that number by the total number of markers. For each data set, the score of the closest marker is compared against all scores of all data sets, so there are 10,100 comparisons per data set. In case of no genetic effect and/or completely random prediction of gene locus, the expected value of this statistic is 0.5. (The fact that all of the 100 data sets are created using the 'true' disease model – instead of the null model – may produce slightly too high estimates of the false-positive rate; however, this is not a problem for purposes of comparison, like in our case.)

POW The power of the method to localize within 5 cM of the true gene. The statistic is the proportion of data sets, where the localization error is lower than or equal to 5 cM. In absence of a genetic effect and/or random guessing, the expected value of the statistic is slightly less than 0.1.

By comparing the values of these statistics, it is possible to draw conclusions about the relative performance of the methods, and also the effect of demographic factors and disease models.

In all experiments, the setup was standardized by fixing a few parameters. The prevalence of the disease was set at 5 per cent. This was incorporated in the disease model, and will be discussed shortly. Second, the current population size was fixed at 100,000 individuals, We focused on a genetic region of 100 cM and had full and error-free genotypes of the affected sample members and their parents. Recombination events were simulated in accordance with Haldane's model, which assumes no chiasma interference.

For determination of the phenotype, we used liability-based models (for exact description, see Section 4.4), where the probability of becoming affected depends on two factors: the presence of a disease mutation and a normally distributed random component. The random component contains factors such as environmental effects and effects of other, unknown genes. The liability of an individual (L) is defined as $L = ax_1 + bx_2 + C$, where indicator variable x_1 indicates the presence of the disease mutation, and x_2 is randomly sampled from standard normal distribution. Constant C is defined in such a way that the desired population prevalence of 5 per cent is reached. For details of the phenotype definition procedure, see Section 4.4. In these experiments, we used three disease models, denoted here by EASY, MEDIUM and HARD. In these models, the relative magnitudes of genetic and random components differ. In model EASY, the ratio $\frac{a}{b}$, corresponding to the relative magnitude of the genetic and environmental component, is set to $\frac{a}{b} = \frac{5}{3}$, in model MEDIUM the ratio is $\frac{a}{b} = 1$, and in the most difficult model HARD, the ratio is $\frac{a}{b} = \frac{2}{3}$.

For simplicity, we utilized a homogeneous marker map, and microsatellite markers with four alleles in each. One of the alleles was always twice as frequent as the others. The PIC value of each marker was 0.678, and the markers were evenly spaced at intervals of 1, 2, and 5 centiMorgans.

8.2 Effect of population history

In this section, we analyze the effect of events in population history such as population substructure, migration, size and shape of the population (pedigree). The population size was fixed at 20 generations, and the number of initial founders was set to 200. Six different founder mutations were introduced to the population.

8.2.1 Population substructure and migration

The population was divided into different subpopulations in the form of a grid: 1×1 (homogeneous population), 2×1, 3×1, 5×1, 2×2, and 3×3. Notation 2×2, for instance, indicates that there were 4 subpopulations altogether organized in the form of a 2×2 grid. Initially, the internal migration rate was set to 0.04. The performance of the methods is shown in Figures 8.1 and 8.2.

With the easiest disease model (EASY), the error and false-positive rates are small and the power adequate. It can be seen that the differences between the methods are small, and the overall effect of substructure to the localization accuracy is low. The accuracy seems to decrease slightly as

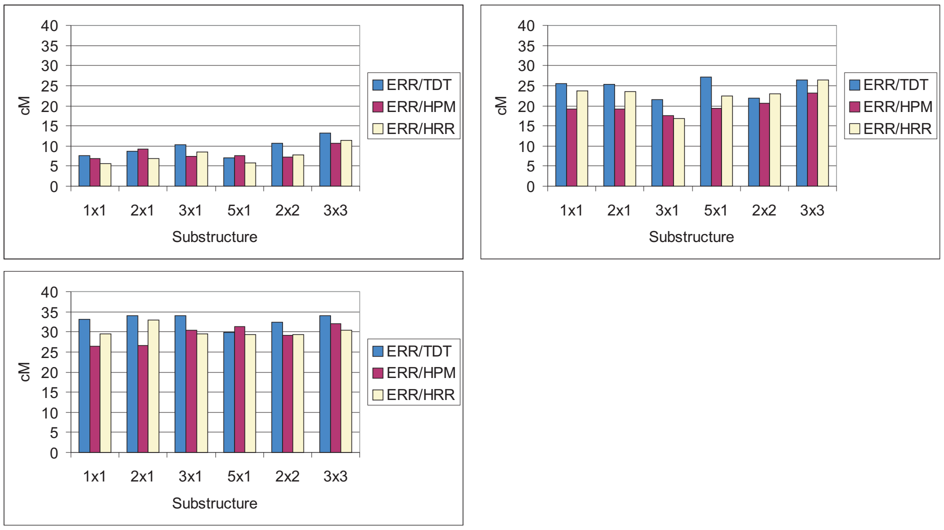

Figure 8.1: Effect of population substructure on localization performance measured by average localization error (ERR). The top left figure corresponds to disease model EASY, the one on the right to model MEDIUM, and the one at the bottom to disease model HARD. In each group of three bars, the bar on the left shows the average error of (multipoint) TDT, the bar in the middle that of Haplotype pattern mining (HPM), and the bar on the right the average localization error of Haplotype relative risk (HRR) approach.

the number of subpopulations increases, but due to the limited number of iterations one must be cautious. When shifting to disease model MEDIUM, the localization errors increase as expected, but no systematic differences between the methods can be seen. As expected, all methods perform poorly with the most difficult disease model (HARD). For the easiest disease model the power to localize the gene at a distance not higher than 5 cM from the true locus is around 0.8 for all population models and methods, and decreases in a systematic way as the disease model becomes more difficult.

In these experiments, the migration rate between subpopulations has been rather high (4 per cent in a generation). Next, it is interesting to see, whether a drop in this rate brings any change in the localization power. For this purpose, the population substructure is fixed to a 3×3 grid, and three different levels of migration are used, namely 0.04, 0.01, and zero. The results are shown in Figure 8.3.

Thus, the effects of both population substructure and population subdivision seem small. This may be due to the robustness of the analysis

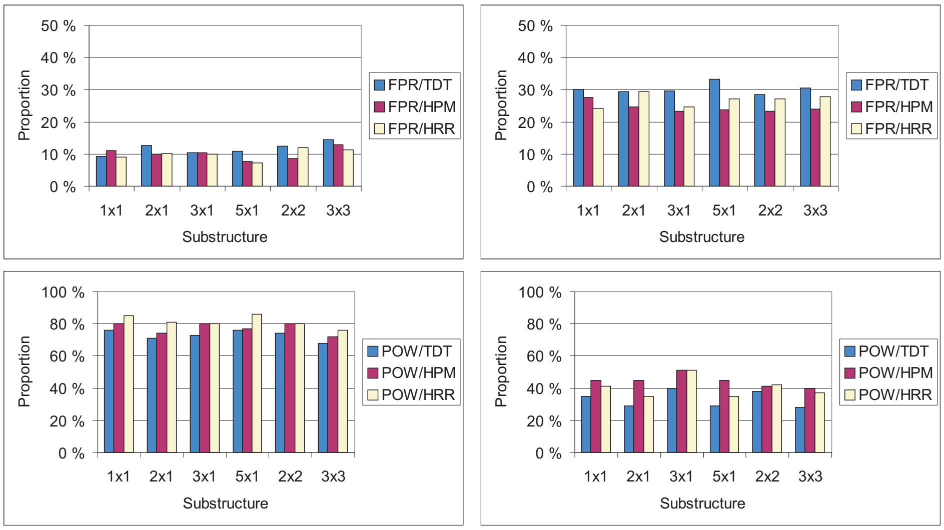

Figure 8.2: Effect of population substructure on localization performance measured by false-positive rate (FPR, top row) and power to localize within 5 cM (POW, bottom row). The two figures on the left correspond to disease model EASY, and the two on the right to model MEDIUM.

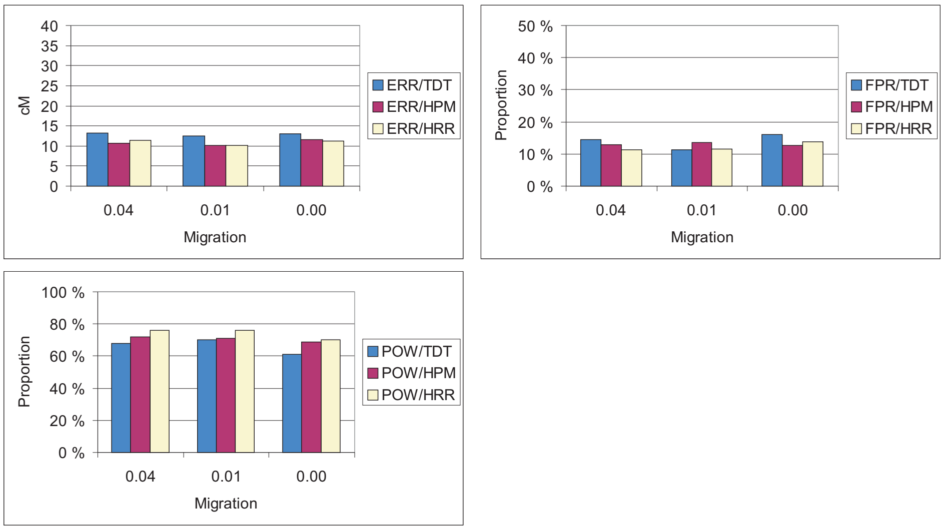

Figure 8.3: Effect of intra-population migration with disease model EASY. The top left figure shows the average localization error (ERR), the one on the right the false-positive rate (FPR), and the one at the bottom the power to localize within 5 cM (POW).

methods in setups where family-based pseudocontrol chromosomes are used. This hypothesis will be tested in Section 8.4.3.

8.2.2 Size and shape of population pedigree

In this section, we change factors affecting the size and shape of the population. These factors are the number of initial founder mutations and the number of generations in the population history.

The size of the initial population changes from 100, 200, and 500 individuals. The number of different founder mutations introduced stays at a constant value of 6, so increasing the number of founder individuals results in a lower proportion of mutation carriers in the population, and, since the prevalence is constant, more phenocopies in the samples. The effect of founder population size on localization power is shown in Figure 8.4.

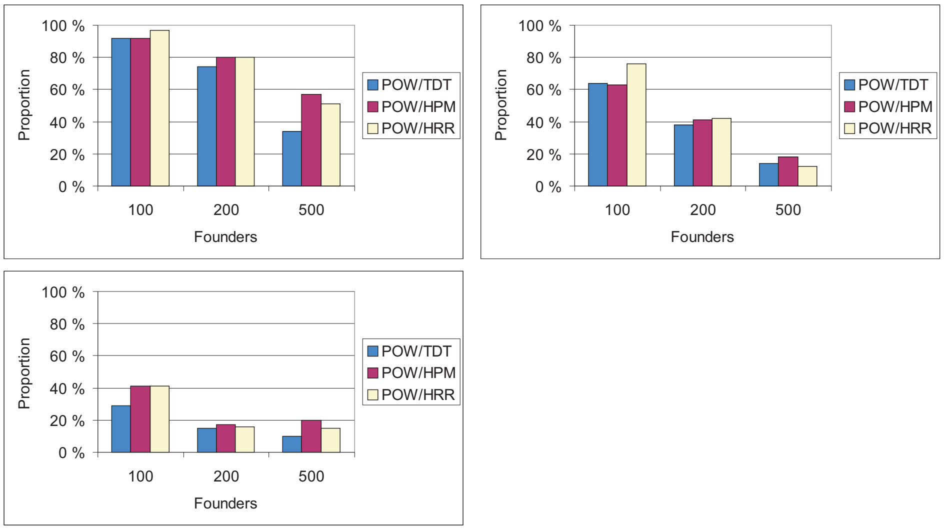

Figure 8.4: Effect of initial population size with disease models EASY (top left), MEDIUM (on the right), and HARD (at the bottom). The localization performance is measured by the power to localize within 5 cM (POW).

Besides the founder population size, the shape of the population pedigree is also controlled by the number of generations since the introduction of the founder population. Figure 8.5 demonstrates the change in localization power, when the number of generations grows from 15 to 20, 25, 30, and, finally, to 50.

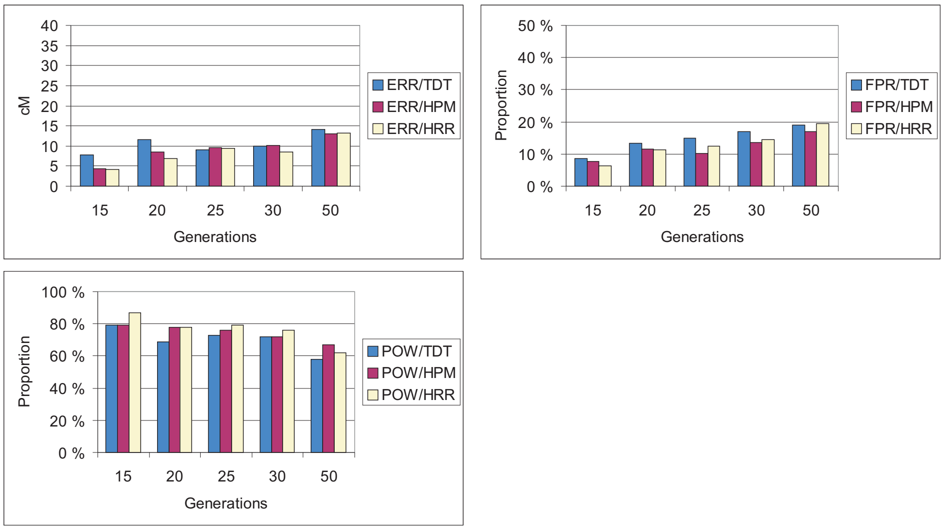

Figure 8.5: Effect of number of generations with disease model EASY. The top left figure shows the average localization error (ERR), the one on the right the false-positive rate (FPR), and the one at the bottom the power to localize within 5 cM (POW).

8.3 Effect of number of founder mutations

Figure 8.6 illustrates the effect of increasing the number of founder mutations increases from 4 to 6, 8, and 10. The power for accurate localization increases, which is due to the design of the simulation procedure: more copies of the original disease allele results in more mutation carriers in the final population. Since the prevalence of the disease is fixed to a high value of 5 per cent, more individuals having a genetic risk factor are available and can end up in the samples. When the number of mutations increases, there are no systematic differences between the methods, except for the harder models, where HPM seems to perform slightly better than the other two methods. In these cases, however, the overall power is low.

8.4 Effect of study design

8.4.1 Marker map

Increasing the distance between adjacent markers should make it more difficult to observe shared haplotypes among individuals, and, thus, deteriorate the power of linkage disequilibrium-based mapping methods. On the other

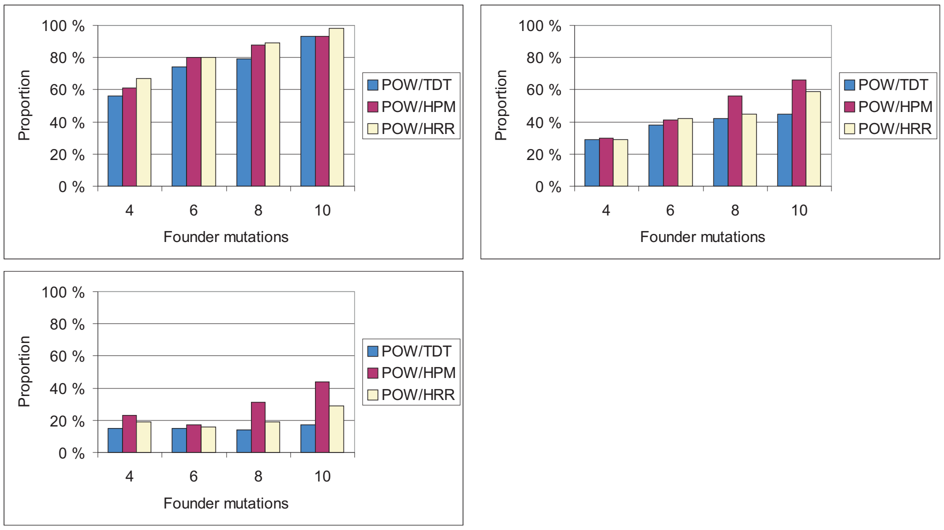

Figure 8.6: Effect of the number of founder mutation with disease models EASY (top left), MEDIUM (right), and HARD (bottom). The localization performance is measured by the power to localize within 5 cM (POW).

hand, using sparse marker maps, especially in the first phase of a genome scan, means savings in genotyping costs. The effect of marker density on the average localization error is shown in Figure 8.7, suggesting that for the selected populations the false-positive rate increases rapidly as the marker map gets more sparse. All three methods are affected equally by this deterioration in performance.

8.4.2 Method of linkage disequilibrium mapping

One of the main messages of the previous sections was that the differences between the three methods analyzed seemed small. In any single setup, the limited number of iterations made it difficult to draw conclusions about their relative performance. However, when all setups are lumped together, and compared pairwise, some differences become visible (Figure 8.8).

First, the performance of HPM is superior to that of TDT in most setups. When compared against HRR, HPM performs better in the majority of cases, but there is a cluster of setups, where the power of HRR is better. These setups seem to be those with high power, indicating that HRR may be advantageous in some easier situations. When HRR in turn was compared against TDT, the former outperforms the latter. The most common order of performance for the three methods was (HPM, HRR, TDT), which was

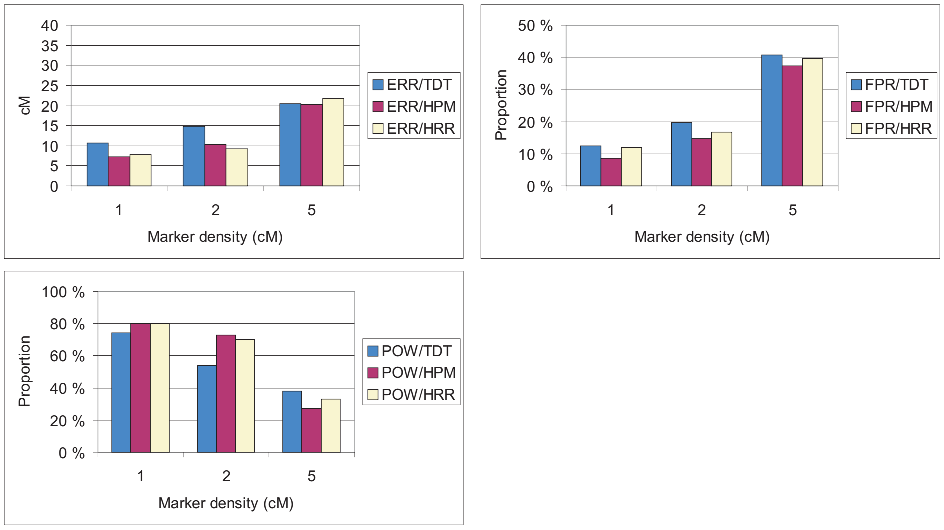

Figure 8.7: Effect of marker map density (1, 2, or 5 cM) with disease model EASY on the false positive rate. The top left figure shows the average localization error (ERR), the one on the right the false-positive rate (FPR), and the one at the bottom the power to localize within 5 cM (POW).

observed in 74/198 cases, and in 18/35 "easy" cases, where the power of at least one method was above 0.5. In these figures, cases where the order of the methods could not be unequivocally defined were omitted. When measured by average localization error, the most common order was again (HPM, HRR, TDT), appearing 126/269 times, and 25/40 times in the easy setup of localization error below 15 cM for at least one of the methods. Finally, when a false-positive rate was used as the measure of performance, the most common order was the same, and it appeared 121/261 times, and 19/38 times in the easy cases, where the rate was below 0.2 for at least one method.

8.4.3 Effect of sampling strategy: family-based vs. population-based control chromosomes

In the previous examples, control chromosomes were created from nontransmitted parental chromosomes. In case-control setups, this is a feasible approach for at least two reasons: firstly, collecting affected individuals and their parents allows deduction of haplotypes, which allows merging information from adjacent markers. Secondly, random individuals and their parental controls tend to share common environmental factors, at least to

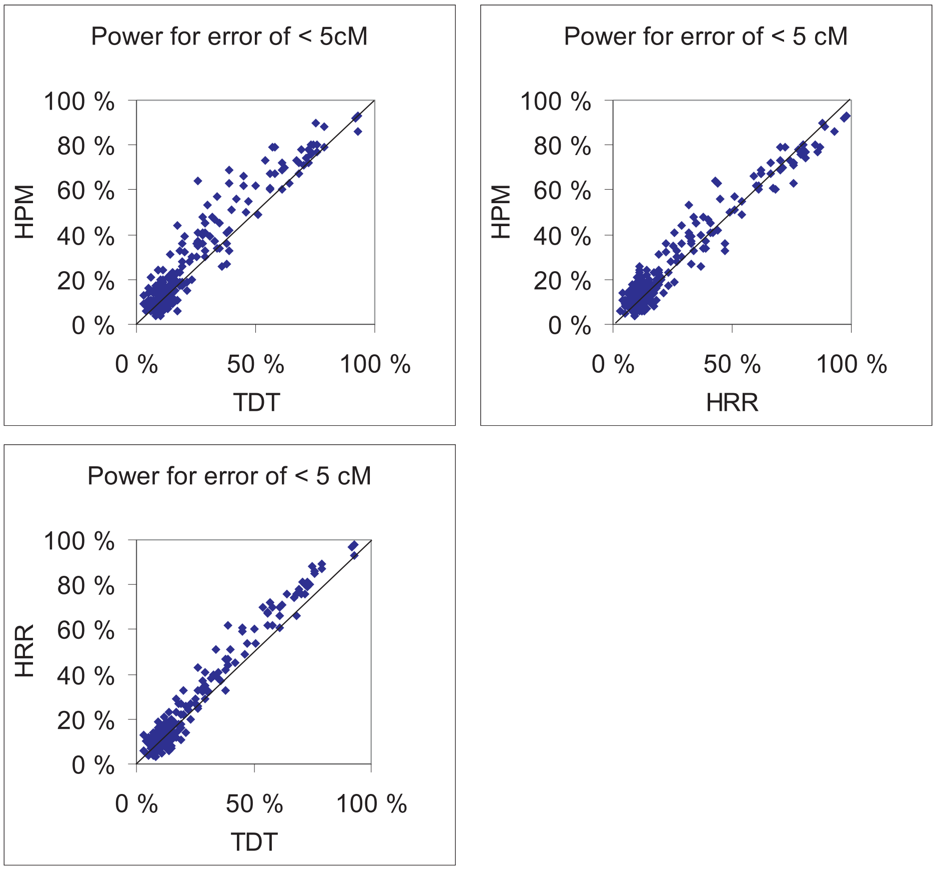

Figure 8.8: Localization methods compared pairwise. The performance is measured by the power to localize within 5 cM (POW). In the top left figure, the power of TDT (horizontal axis) is compared to that of HPM (vertical axis). On the right, the comparison is made between HRR and HPM, and at the bottom between TDT and HRR. Each dot in the figures corresponds to a single setup (consisting of a population model, number of founder mutations, and map density).

some extent.

If trios are not collected, an alternative approach is to use random unaffected individuals collected from the entire population. The tests of allelic association may, however, be dangerous due to hidden population stratification, as will be shown in the following results.

The effect of the sampling strategy was analyzed with a population that had a 3×3 substructure, and a migration rate of 0.01 between the subpopulations. Each simulation of 100 data sets was repeated 10 times from the beginning, i.e. creation of the population, yielding a total of 1,000 data sets, which are first analyzed jointly. The results are shown in figure 8.9.

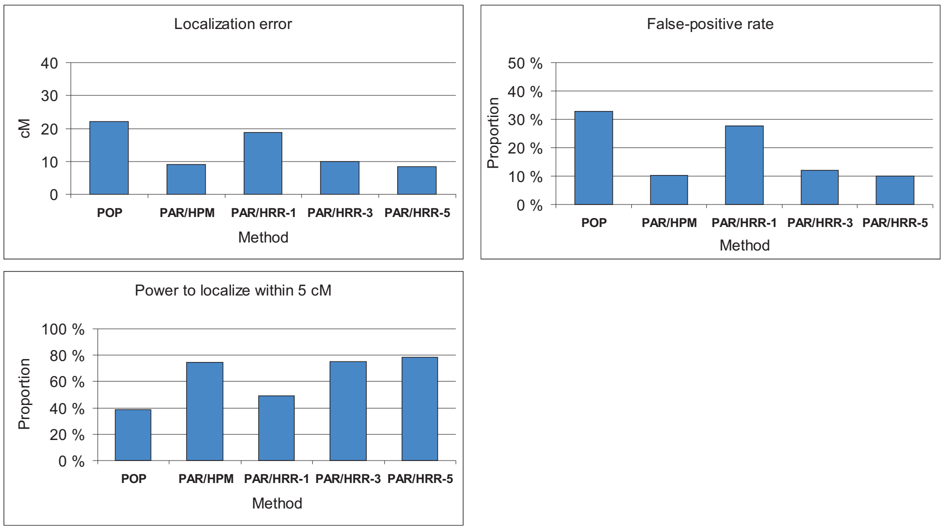

Figure 8.9: Effect of sampling strategy and haplotype length for disease model EASY and 3×3 substructure. The top left figure shows the average localization error (ERR), the one on the right the false-positive rate (FPR), and the one at the bottom the power to localize within 5 cM (POW). In all three figures, the methods compared are haplotype relative risk with population-based controls (POP), HPM with parental controls (PAR/HPM), and HRR with maximum haplotype length of 1, 3, and 5 adjacent markers (PAR/HRR-1, PAR/HRR-3, and PAR/HRR-5).

Next, the variation between the 10 collections of 100 data sets is analyzed. For each of the 10 collections, the entire population (pedigree) is simulated from scratch. The purpose of the experiment is to verify some of the earlier results: if there is not much variation between the groups of 100 data sets – each corresponding to one simulation of the population and inheritance of chromosomal segments – then the results can be considered reliable with respect to the small number of iterations. The results shown in Figure 8.10 indicate that the variation is relatively small, but in order to make profound predictions about the exact performance of any single study design, a higher number of iterations may be desirable.

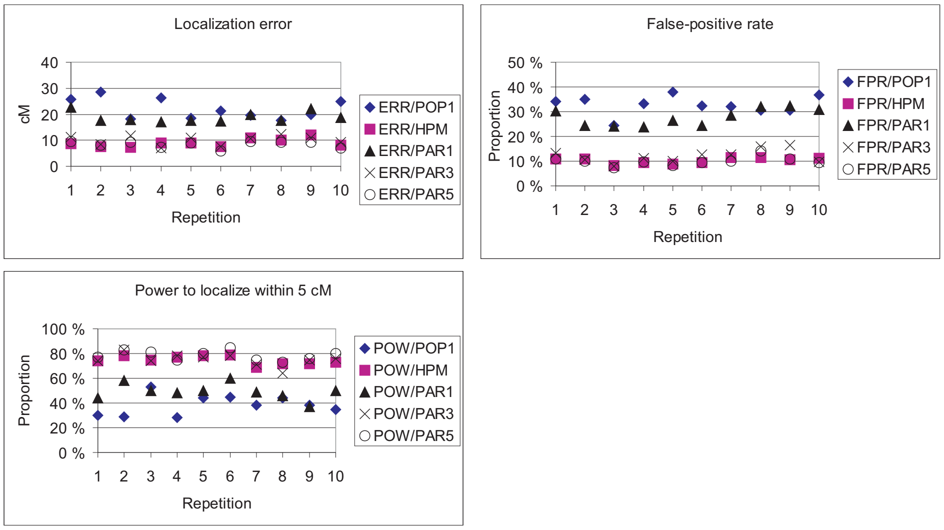

Figure 8.10: Variation between collections of 100 data sets (numbered $1, 2, \ldots, 10$), where each collection is re-simulated starting from the construction of the population (pedigree). The disease model was EASY and the population had a 3×3 substructure. The top left figure shows the average localization error (ERR), the one on the right the false-positive rate (FPR), and the one in the bottom the power to localize within 5 cM (POW). In all figures, the methods compared are haplotype relative risk with population-based controls (POP), HPM with parental controls (PAR/HPM), and HRR with maximum haplotype length of 1, 3, and 5 adjacent markers (PAR/HRR-1, PAR/HRR-3, and PAR/HRR-5).

The problem with using population-based pseudocontrols in the presence of population substructure seems to be the fact that if a considerable amount of mutation-carriers comes from the same subpopulation(s), then several – or even all – marker alleles common to that subpopulation may become strongly associated to the disease phenotype. Unfortunately, these spurious associations can be observed anywhere in the genome, irrespective of the chromosomal location of the disease gene.

Chapter 9

Challenge of multifactorial diseases

The etiology of common human diseases is a complex issue. The probability of an individual to become affected — or the severity of the phenotype of an individual — may depend on complex interactions between multiple genes and the environment. The complexity of this process is also a challenge for disease gene localization, and should be taken into account at the very beginning of the disease gene localization project. One of the major findings in an extensive meta-analysis on 101 genome scans for multifactorial diseases [APF01] was that there is no uniformly best strategy for mapping human complex diseases. According to [APF01], increased sample size and reduced genetic heterogeneity yielded a higher success rate in a mapping project. On the other hand, the nature of phenotype (qualitative vs. quantitative), prevalence of disease, degree of isolation in population history and recurrence risk (probability that a relative of given kinship is affected conditional of an affected index case) did not turn out to be significant determinants for the success of a genome scan. In light of this finding, the importance of sufficient descriptive analysis in the early stage of a project is stressed: the relationship of the phenotypes and covariates to the sampling strategy, study design and genetic analysis method should be thoroughly analyzed [APF01].

The purpose of the following experiment is to illustrate the relationship between the effect of a disease gene and the applicability of association-based methods.

Let us consider a common multifactorial disease, whose population prevalence is 0.04. The disease is going to be studied using a case/control setup with an equal number of individuals in both groups. We focus on the ideal case, where a fully informative marker is located exactly on top of the disease locus, and compute the power for obtaining a significant result in the association test computed from a 2×2 contingency table. We ignore the effect of multiple testing, a factor that should be considered in setting critical thresholds in real analysis. Thus, the results obtained will provide an upper limit of what can be seen in reality.

The effect of the disease-predisposing allele is expressed by two penetrances $p_{A|R}$ and $p_{A|N}$, indicating probabilities for becoming affected for carriers and noncarriers of the predisposing allele. In the simulations, samples are created by deciding independently for each individual in either sample, whether he/she is a carrier of the risk allele. The decision is based on conditional probabilities $p_{R|A} = \frac{p_{A|R} \cdot p_R}{p_A}$, $p_{N|A} = 1 - p_{R|A}$, $p_{R|C} = \frac{p_{C|R} \cdot p_R}{p_C}$, and $p_{N|C} = 1 - p_{R|C}$, where R and N indicate carrying and not carrying the disease-predisposing allele, and A and C indicate affected and control individuals. Constant p_R is the prevalence of the predisposing allele, which can be expressed as $p_R = \frac{p_A - p_{A|1}}{p_{A|2} - p_{A|1}}$ where p_A indicates the prevalence of the disease in the population.

In the simulations, we consider sample sizes of 20, 50, 100, and 200 affected individuals plus a similar number of healthy controls. For each sample size and disease model, 10,000 replicates are generated. Within each replicate, affected and control samples are constructed, and the counts for mutation carriers and non-carriers in both samples are determined. The χ^2 test of association is then done for these counts, and it is checked whether the critical thresholds corresponding to p values of 0.05, 0.01, and 0.001 are exceeded. The proportion of such replicates, where a given threshold was exceeded is the power to detect association for the desired disease model. Different disease models are tested by letting the penetrance of non-carriers ($p_{A|N}$) vary between 0 and 0.04 (a higher value is impossible because of the selected population prevalence), whereas the penetrance of carriers varies between zero and unity. In addition, it is required that each pair of penetrances must produce a disease carrier frequency between 0 and 1. The relationship of penetrances and disease carrier frequency is shown in Figure 9.1.

The results of the simulations are depicted in Figures 9.2 and 9.3. It can be seen that even with relatively small sample sizes, the power to detect association is very high, if either the penetrance of non-carriers is low or the penetrance of carriers is high. As expected, increasing the sample size helps, but, there is always the area of low power, when penetrances of non-carriers are close to penetrances of carriers. For instance, for the largest sample size of 200 affected and 200 control individuals, there is a large family of genetic models, where the power to detect association at the level of 0.01 is below 70 per cent (Table 9.1). It can be seen that there is a wide variety of true disease models where the power is poor: for these models, the relative

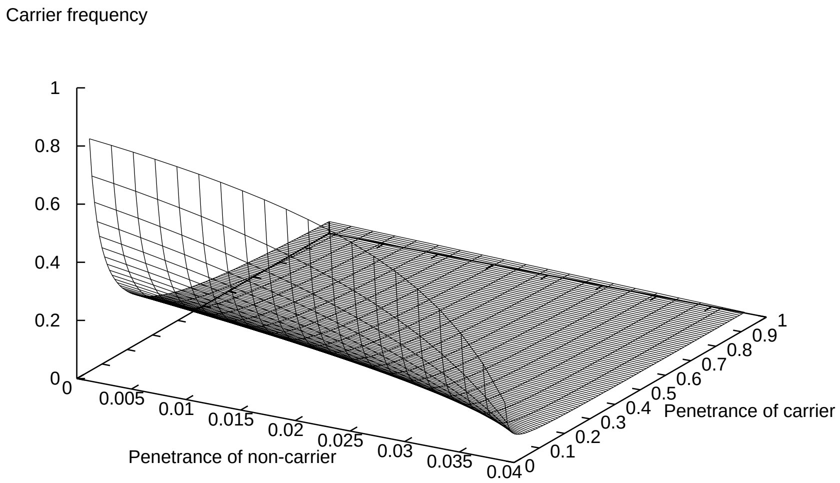

Figure 9.1: Carrier frequency as a function of penetrances, assuming fixed prevalence of 0.04. The axes in the bottom correspond to penetrances of non-carriers and carriers, respectively. The vertical axis shows the carrier frequency required to explain the selected penetrance of 0.04.

risks vary between 1.25 and 7.5, and the frequency of the disease allele can be practically anything. Thus, it is not surprising that the relative risk attributed to the disease locus is not a good predictor of the localization success, as pointed out in [APF01].

Like most methods described earlier in this work, most disease gene localization methods try to detect one disease-predisposing gene at a time. Even if there were interaction between several genes, each of which contributes to the disease status, the signals of the genes may in some cases be strong enough to detect individually. The likelihood-based method of Cordell et al. [CWJ00] has similar power for independent searches of one gene at a time, if the disease model is multiplicative, but offers increased power in other cases. In the method, the original ASP statistic (see page 53) is extended to multiple loci. The significance of the observed statistic value can be obtained by simulation, but it is pointed out that instead of stringent significance criteria one may be interested in getting a general

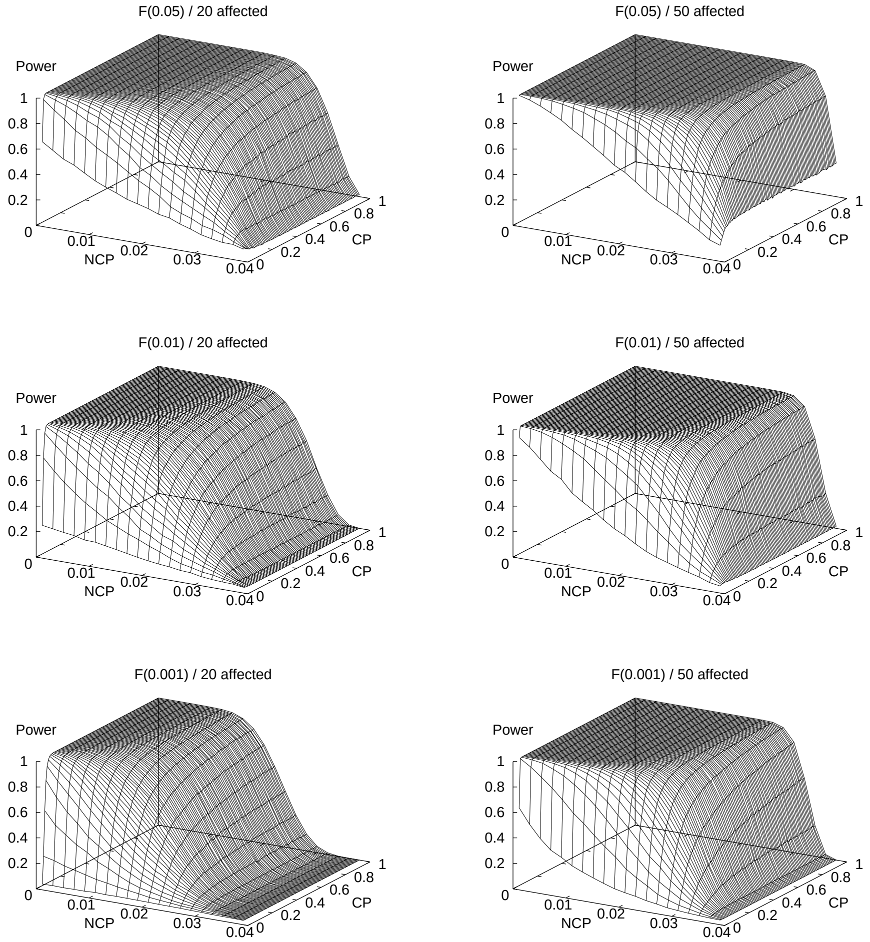

Figure 9.2: Power to detect association as a function of the true disease model, sample size and the significance level. There are two sample sizes, 20 (on the left) and 50 (on the right) and three significance levels, 0.05 (at the top), 0.01 (in the middle), and 0.001 (at the bottom). In each figure, the penetrances on non-carriers (NCP) and carriers (CP) are in the horizontal axes, and the corresponding power is on the vertical axis.

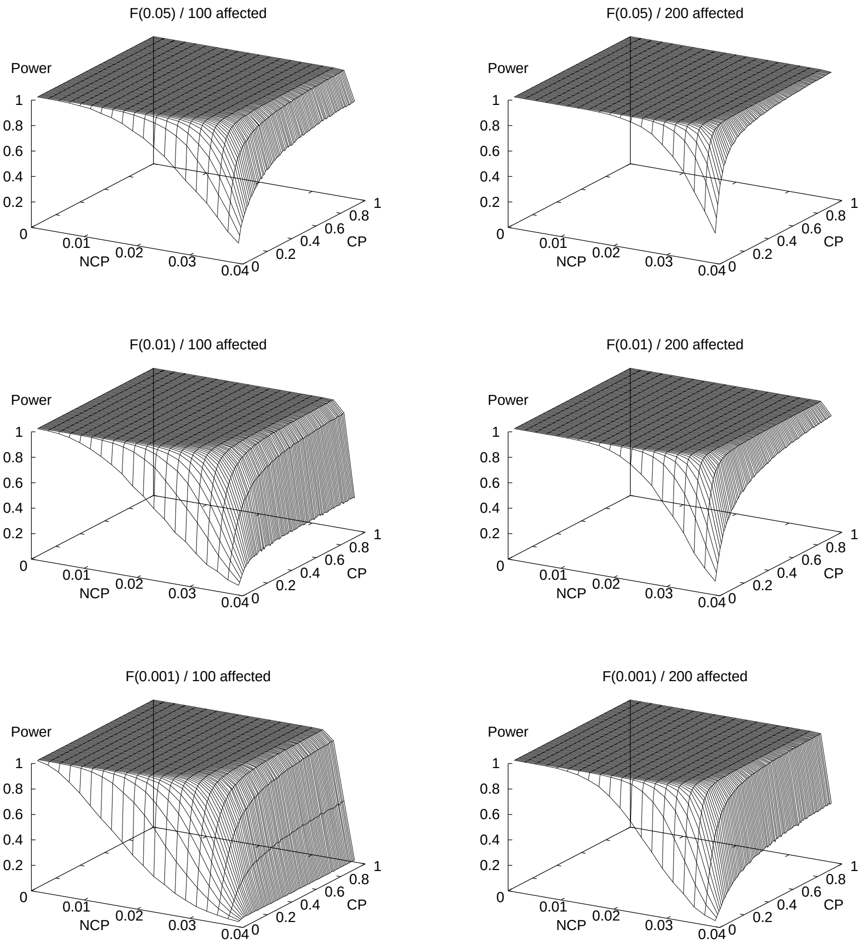

Figure 9.3: Power to detect association as a function of the true disease model, sample size and the significance level. There are two sample sizes, 100 (on the left) and 200 (on the right) and three significance levels, 0.05 (at the top), 0.01 (in the middle), and 0.001 (at the bottom). In each figure, the penetrances on non-carriers (NCP) and carriers (CP) are in the horizontal axes, and the corresponding power is on the vertical axis.

$p_{A\|N}$	$p_{A\|R}$	p_R	F(0.05)	F(0.01)	F(0.001)
0.0280	0.0500	0.5455	0.8260	0.6273	0.3523
0.0300	0.0500	0.5000	0.7468	0.5260	0.2584
0.0320	0.0500	0.4444	0.6435	0.4123	0.1813
0.0340	0.0500	0.3750	0.5011	0.2684	0.0928
0.0340	0.0600	0.2308	0.7639	0.5361	0.2631
0.0360	0.0500	0.2857	0.3612	0.1626	0.0431
0.0360	0.0600	0.1667	0.5863	0.3354	0.1207
0.0360	0.0700	0.1176	0.7228	0.4856	0.2087
0.0360	0.0800	0.0909	0.8015	0.5888	0.2985
0.0360	0.0900	0.0741	0.8537	0.6699	0.3699
0.0380	0.0500	0.1667	0.2035	0.0747	0.0140
0.0380	0.0600	0.0909	0.3173	0.1370	0.0343
0.0380	0.0700	0.0625	0.4162	0.2019	0.0530
0.0380	0.0800	0.0476	0.4920	0.2572	0.0724
0.0380	0.0900	0.0385	0.5653	0.3138	0.0873
0.0380	0.1000	0.0323	0.6055	0.3475	0.1066
0.0380	0.1100	0.0278	0.6462	0.3915	0.1274
0.0380	0.1200	0.0244	0.7002	0.4298	0.1467
0.0380	0.1300	0.0217	0.7126	0.4526	0.1568
0.0380	0.1400	0.0196	0.7435	0.4797	0.1757
0.0380	0.1500	0.0179	0.7662	0.5008	0.1913
0.0380	0.1600	0.0164	0.7804	0.5133	0.1961
0.0380	0.1700	0.0152	0.7978	0.5332	0.2150
0.0380	0.1800	0.0141	0.8197	0.5628	0.2231
0.0380	0.1900	0.0132	0.8276	0.5700	0.2320
0.0380	0.2000	0.0123	0.8318	0.5808	0.2478
0.0380	0.2100	0.0116	0.8418	0.5966	0.2528
0.0380	0.2200	0.0110	0.8591	0.6074	0.2651
0.0380	0.2300	0.0104	0.8714	0.6269	0.2663
0.0380	0.2400	0.0099	0.8709	0.6325	0.2794
0.0380	0.2500	0.0094	0.8810	0.6524	0.2908
0.0380	0.2600	0.0090	0.8871	0.6616	0.2963
0.0380	0.2700	0.0086	0.8921	0.6729	0.2962
0.0380	0.2800	0.0083	0.8949	0.6666	0.2987
0.0380	0.2900	0.0079	0.8963	0.6790	0.3080
0.0380	0.3000	0.0076	0.9064	0.6924	0.3239

Table 9.1: Some disease models with low power, i.e. $F(0.01) < 0.7$, computed from samples of 200 affected and 200 control individuals. The columns from left to right are: penetrances of the non-carriers and carriers of the disease-predisposing mutation, frequency of the risk allele carriers required to explain the assumed prevalence of 0.04, and estimated power for significance levels of 0.05, 0.01, and 0.001.

pattern of results and finding regions that provide a starting point for further investigation [CWJ00]. Strauch et al. [SFK00] present an extension of parametric and nonparametric linkage analysis to two trait loci, located in two distinct marker maps that are not genetically linked. The approach is implemented in GENEHUNTER-TWOLOCUS program. In the implementation, the scoring functions (NPL statistics) in the nonparametric analysis are simply sums of the scoring functions for the two independent loci. In the field of data-mining, multilocus localization is also feasible. In [TOV00b], two-chromosome patterns were successfully searched for from two-locus data sets created by the population simulator described in section 4.

The limits of localization capability can be pushed further by paying close attention to the phenotype. Whenever the phenotype studied is by nature quantitative (such as blood pressure, or concentration of a substance in blood), quantitative linkage analysis may be the natural choice. The possible benefits include more precise definition of the phenotype for affected individuals and the possibility to use an intermediate phenotype, which may be affected by fewer genetic factors than the disease trait [WeL95].

Chapter 10

Conclusions

In this work we created a method and software for simulating the population history of an isolate, inheritance of chromosomes within it as well as determining the phenotypes and sampling of individuals, and evaluation of the power of the intended study strategy. Also, a method for finding the optimal strategy for a multistage genome scan study was introduced. Because of the complex nature of the phenomena of interest, the experiments relied heavily on simulation techniques.

The selection of an empirical approach was, however, deliberate. The field of computational challenges in disease gene localization is getting wider due to technological advances and interest towards describing more complex relationships between the phenotypes and the underlying genes. This may increase the demand for simulation techniques, where rigid statistical assumptions are not incorporated to a similar extent as in the analytical approach.

New biotechnical inventions have created a situation, where large amounts of genetic data can easily be obtained. Genomewide maps of polymorphic markers are nowadays widely used, and very dense maps of single nucleotide polymorphisms are soon commercially available and economically feasible. Even though the limited information content of marker maps will not be a problem, this may create another drawback: when the amount of information increases manifold, it is not sufficient to simply increase the number of hypotheses tested: new thinking and new algorithmic approaches are needed to deal with problems of multiple testing.

The advantages made in the marker biochemistry are further challenged by common, multifactorial diseases, where the relationship between several disease-predisposing genes and environmental effects is complex. Traditional parametric linkage analysis, which relies on assumptions about the inheritance model of the disease often results in very poor power, which has created a demand for alternative approaches. One such approach is genomewide haplotype analysis in genetic isolates, where the sample members are assumed to contain less genetic heterogeneity in the disease loci compared to random disease families sampled from the general population. The decreased level of genetic heterogeneity can result in easier localization of the underlying disease genes. The methods for this kind of analysis have increased during the last few years: there are approaches that utilize multiple adjacent markers, allow for several founder chromosomes or even multiple disease genes simultaneously. Some are based on complex likelihood functions or advanced data mining approaches. All of these methods have their own features and drawbacks: some try to model the coalescence history of the disease chromosomes very accurately, while others consider it as a 'black box'. Obviously, there is a lot of choice, but one may ask, how fruitful it really is to go further in the fine-tuning methods of association analysis, unless new information about the population history and disease phenotype is actively incorporated and used in the analysis procedure: some of the methods may already be at the level where only marginal improvements in the localization power are possible, if the information content of the input does not increase. In critical evaluation of different localization methods, it is vital to compare the localization performance using a large number of simulated data sets.

The computational problems in localization of complex diseases go far beyond the selection of the optimal analysis methods. Some factors in population history (number of generations since the mutation was introduced, number of founders, number of copies of the mutation allele) have a considerable effect on what can be seen from the data. These factors must be taken into account at the early stages of the mapping project, when decisions on the sampling strategy and sample size are made. Even though it may be inaccurate to assess the power of any specific project in advance by incorporating all information about the population history into a single run of a simulation program, it is still a highly recommendable idea to perform pre-analysis *in silico*. Intelligent guesses can be made about the bounds of the population parameters and the disease model. By simulating the actual, intended analysis procedure with data sets generated under various assumptions, one can find the area of parameters where the mapping project is feasible as well as the area, where it is not. This information may be very helpful in considering, whether to commence the project. Furthermore, it may help in finding the optimal strategy, which includes selection of the density of the marker map in the first phase, as well as the required sample size. Eventually, the selection of the optimal strategy is a matter of minimizing the cost of the study, a framework for which was presented in Section 5.2.

One of the computational challenges in disease gene localization is in the interpretation of the significance of the observed signal, whether it is linkage to a locus or association to the phenotype. In the first phase of the genome scan, however, one is not necessarily interested in a significant result, but the significance should rather be used in sorting the areas in the genome into decreasing order of interest.

Finally, there are new challenges that probably become more critical in the near future. In pharmacogenetics studies the focus of interest is the relationship between drug response and genetic factors, and demand for computational methods that can handle this kind of data is almost certainly going to increase, when more information about the loci and functions of disease-predisposing genes becomes available.

References

[AMS96] R. Agrawal, H. Mannila, R. Srikant, H. Toivonen and A. I. Verkamo, Fast discovery of association rules. In U. M. Fayyad, G. Piatetsky-Shapiro, P. Smyth and R. Uthurusamy (eds.), Advances in Knowledge Discovery and Data Mining, AAAI Press, 1996, 307-328.

[APF01] J. Altmüller, L. J. Palmer, G. Fischer, H. Scherb and M. Wjst, Review article: Genomewide scancs of complex human diseases: true linkage is hard to find. *American Journal of Human Genetics* **69** (2001), 936-950.

[Bal01] F. Balloux, EASYPOP (Version 1.7): A computer program for population genetics simulations. *The Journal of Heredity* **92** (2001), 301-302.

[BGH01] C. Bourgain, E. Génin, P. Holopainen, K. Mustalahti, M. Mäki, J. Partanen and F. Clerget-Darpoux, Use of closely related affected individuals for the genetic study of complex diseases in founder populations. *American Journal of Human Genetics* **68** (2001), 154-159.

[BrE99] T. G. Bruun and H. Ewald, Selection bias of susceptibility genes possible when using parent-offspring trios in genetic association studies. *Molecular Psychiatry* **4** (1999), 414-416.

[BrW00] K. W. Broman and J. Weber, Characterization of human crossover interference. *American Journal of Human Genetics* **66** (2000), 1911-1926.

[ChD94] G. A. Churchill and R. W. Doerge, Empirical threshold values for quantitative trait mapping. *Genetics* **138** (1994), 963-971.

[CWJ00] H. J. Cordell, G. C. Wedig, K. B. Jacobs and R. C. Elston, Multilocus linkage tests based on affected relative pairs. *American Journal of Human Genetics* **66** (2000), 1273-1286.

[Del93] A. de la Chapelle, Disease gene mapping in isolated human populations: the example of Finland. *Journal of Medical Genetics* **30** (1993), 857-865.

[DeR95] B. Devlin and N. Risch, A comparison of linkage disequilibrium measures for fine-scale mapping. *Genomics* **29** (1995), 311-322.

[DeW98] A. de la Chapelle, F. Wright, Linkage disequilibrium mapping in isolated populations: The example of Finland revisited. *Proceedings of the National Academy of Sciences of the United States of America* **95** (1998), 12416-12423.

[DFF96] C. Dib, S. Faure, C. Fizames, D. Samson, N. Drouot, A. Vignal, P. Millasseau, S. Marc, J. Hazan, E. Seboun, M. Lathrop, G. Gyapay, J. Morissette J. and J. Weissenbach, A comprehensive genetic map of the human genome based on 5,264 microsatellites. *Nature* **380** (1996), 152-154.

[DSG96] S. Davis, M. Schroeder, L. R. Goldin and D. E. Weeks, Nonparametric simulation-based statistics for detecting linkage in general pedigrees. *American Journal of Human Genetics* **58** (1996), 867-880.

[DuF97] L. Durham and E. Feingold, Genome scanning for segments shared identical by descent among distant relatives in isolated populations. *American Journal of Human Genetics* **61** (1997), 830-842.

[Els97] R. Elston, 1996 William Allan award address: Algorithms and Inferences: the challenge of multifactorial diseases. *American Journal of Human Genetics* **60** (1997), 255-262.

[FaR87] C. T. Falk and P. Rubinstein, Haplotype relative risks: an easy reliable way to construct a proper control sample for risk calculations. *Annals of Human Genetics* **51** (1987), 227-233.

[Gau95] W. J. Gauderman, A method for simulating familial disease data with variable age at onset and genetic and environmental effects. *Statistics and Computing* **5** (1995), 237-243.

[Goo00] P. Good, Permutation Tests: a Practical Guide to Resampling Methods for Testing Hypotheses, 2nd ed., Springer-Verlag, New York, 2000, 13-16, 24-25, 184-186.

[GrD96] D. A. Greenberg and P. Doneshka, Partitioned association-linkage test: distinguishing "necessary" from "susceptibility" loci. *Genetic Epidemiology* **13** (1996), 243-252.

[Guo97] S.-W. Guo, Linkage disequilibrium measures for fine-scale mapping: a comparison. *Human Heredity* **47** (1997), 301-314.

[HBB94] H. J. Houwen, S. Baharloo, K. Blankenship, P. Raeymaekers, J. Juyn, L. A. Sandkuijl and N. B. Freimer, Genome screening by searching for shared segments: mapping a gene for benign recurrent intrahepatic cholestasis. *Nature Genetics* **8** (1994), 380-386.

[HBG96] E. R. Hauser, M. Boehnk, S.-W. Guo and N. Risch, Affected-sib-pair interval mapping and exclusion for complex genetic traits: sampling considerations. *Genetic Epidemiology* **13** (1996), 117-137.

[HDK92] J. Hästbacka, A. de la Chapelle, I. Kaitila, P. Sistonen, A. Weaver and E. Lander, Linkage disequilibrium mapping in isolated founder populations: diastrophic dysplasia in Finland. *Nature Genetics* **2** (1992), 204-211.

[HVS99] I. Hovatta, T. Varilo, J. Suvisaari, J. Terwilliger, V. Ollikainen, R. Arajärvi, H. Juvonen, M. L. Kokko-Sahin, L- Väisänen, H. Mannila, J. Lönnqvist and L. Peltonen, A genomewide screen for schizophrenia genes in an isolated Finnish subpopulation, suggesting multiple susceptibility loci. *American Journal of Human Genetics* **63** (1999), 1114-1124.

[HWS98] J. Hampe, T. Wienker, S. Schreiber and P. Nurnberg , POPSIM: a general population simulator program. *Bioinformatics* **14** (1998), 458-464.

[JCW95] L. Jorde, J. Carey and R. White, Medical Genetics. Mosby, St. Louis, 1995.

[Joh75] S. Johnson, Yacc: Yet Another Compiler Compiler, Computing Science Technical Report No. 32, Bell Laboratories, Murray Hill, New Jersey, 1975.

[Jor00] L. Jorde, Linkage disequilibrium and the search for complex disease genes. *Genome Research* **10** (2000), 1435-1444.

[KDR96] L. Kruglyak, M. Daly, M. Reeve-Daly, E. Lander, Parametric and nonparametric linkage analysis: a unified multipoint approach. *American Journal of Human Genetics* **58** (1996), 1347-1363.

[KMN94] S. Koskinen, T. Martelin, I.-L. Notkola, V. Notkola and K. Pitkänen, Suomen väestö. Gaudeamus, Hämeenlinna, 1994.

[KSB94] M. Knapp, S. A. Seuchter and M. P. Baur, Linkage analysis in nuclear families 2: relationship between affected sib-pair tests and lod score analysis, *Human Heredity* **44** (1994), 44-51.

[LaK95] E. S. Lander and L. Kruglyak, Genetic dissection of complex traits: guidelines for interpreting and reporting linkage results. *Nature Genetics* **11** (1995), 241-247.

[Les75] M. Lesk, Lex - A Lexical Analyzer Generator, Computing Science Technical Report No. 39, Bell Laboratories, Murray Hill, New Jersey, 1975.

[Lew88] R. C. Lewontin, On measures of gametic disequilibrium. *Genetics* **120** (1988), 849-852.

[LNJ01] B. Leyanaert, C. Neukirch, D. Jarvis, S. Chinn, P. Burney and F. Neukirch, Does living on a farm during childhood protect against asthma, allergic rhinitis, and atopy in adulthood? *American Journal of Respiratory and Critical Care Medicine* **164** (2001), 1829-1834.

[LLB01] E. S. Lander, L. M. Linton, B. Birren, C. Nusbaum, M. C. Zody, J. Baldwin, K. Devon, K. Dewar, M. Doyle, W. FitzHugh, et al., Initial sequencing and analysis of the human genome. *Nature* **409** (2001), 860-921.

[LLJ84] G. M. Lathrop, J. M. Lalouel, C. Julier and J. Ott, Strategies for multilocus linkage analysis in humans. *Proceedings of the National Academy of Sciences of the United States of America* **81** (1984), 3443-3446.

[LoL99] A. D. Long and C. H. Langley, The power of association studies to detect the contribution of candidate genetic loci to variation in complex traits. *Genome Research 9* (1999), 720-731.

[MaT97] H. Mannila and H. Toivonen, Levelwise search and borders of theories in knowledge discovery. *Data Mining and Knowledge Discovery* **1** (1997), 241-258.

[McS99] M. McPeek and A. Strahs, Assessment of linkage disequilibrium by the decay of haplotype sharing, with application to fine-scale genetic mapping. *American Journal of Human Genetics* **65** (1999), 858-875.

[MMW00] E. R. Martin, S. A. Monks, L. L. Warren and N. L. Kaplan, A test for linkage and association in general pedigrees: the pedigree disequilibrium test. *American Journal of Human Genetics* **67** (2000), 146-154.

[Mor55] N. E. Morton, Sequential tests for the detection of linkage. *American Journal of Human Genetics* **7** (1955), 277-318.

[MSK93] C. J. MacLean, P. C. Sham and K. S. Kendler, Joint linkage of multiple loci for a complex disorder. *American Journal of Human Genetics* **53** (1993), 353-366.

[MWB02] A. P. Morris, J. C. Whittaker and D. J. Balding, Fine-scale mapping of disease loci via shattered coalescent modeling of genealogies. *American Journal of Human Genetics* **70** (2002), 686-707.

[NMH01] J. Nopola-Hemmi, B. Myllyluoma, T. Haltia, M. Taipale, V. Ollikainen, T. Ahonen, A. Voutilainen, J. Kere, E. Widén, A dominant gene for developmental dyslexia on chromosome 3. *Journal of Medical Genetics* **38** (2001), 658-664.

[OtR97] J. Ott and D. Rabinowitz, The effect of marker heterozygosity on the power to detect linkage disequilibrium. *Genetics* **147** (1997), 927-930.

[Ott89] J. Ott, Computer-simulation methods in human linkage analysis. *Proceedings of the National Academy of Sciences of the United States of America* **86** (1989), 4175-4178.

[Ott99] J. Ott, Analysis of Human Genetic Linkage. The Johns Hopkins University Press, Baltimore, 1999.

[OOS02] P. Onkamo, V. Ollikainen, P. Sevon, H. T. T. Toivonen, H. Mannila and J. Kere, Association analysis for quantitative traits by data mining: QHPM. *Annals of Human Genetics*. In press.

[Per98] T. V. Perneger, What's wrong with Bonferroni adjustments. *British Medical Journal* **316** (1998), 1236-1238.

[Pri01] J. K. Pritchard, Are rare variants responsible for susceptibility for complex diseases? *American Journal of Human Genetics* **69** (2001), 124-137.

[Ris90] N. Risch, Linkage strategies for genetically complex traits: I. multilocus models. *American Journal of Human Genetics* **46**(1990), 222-228.

[Ris90b] N. Risch, Linkage strategies for genetically complex traits. II.

The power of affected relative pairs. *American Journal of Human Genetics* **46**(1990), 229-241.

[RWC81] P. Rubinstein, M. Walker, C, Carpenter, C. Carrier, J. Krassner, C. Falk, F. and Ginsberg, F, Genetics of HLA disease associations: the use of the haplotype relative risk (HRR) and the haplo-delta (Dh) estimates in juvenile diabetes from racial groups (abstract). *Human Immunology* **3** (1981), 384.

[SFK00] K. Strauch, R. Fimmers, T. Kurz, K. Deichmann, T. Wienker and M. Baur, Parametric and nonparametric multipoint linkage analysis with imprinting and two-locus-trait models: application to mite sensitization. *American Journal of Human Genetics* **66** (2000), 1945-1957.

[SGE93] R. Spielman, R. McGinnis and J. Warren, Transmission test for linkage disequilibrium: the insulin gene region and insulin-dependent diabetes mellitus (IDDM). *American Journal of Human Genetics* **52** (1993), 506-516.

[SLF99] S. K. Service, D. W. T. Lang, N. Freimer and L. A. Sandkuijl, Linkage-disequilibrium mapping of disease genes by reconstruction of ancestral haplotypes of founder populations. *American Journal of Human Genetics* **64** (1999), 1728-1738.

[SRR78] B. Suarez, J. Rice and T. Reich, The generalized sib pair IBD distribution: its use in the detection of linkage. *Annals of Human Genetics* **42** (1978), 87-94.

[SRT02] D. J. Schaid, C. M. Rowland, D. E. Tines, R. M. Jacobson and G. A. Poland, Score tests for association between trais and haplotypes when linkage phase is ambiguous. *American Journal of Human Genetics* **70** (2002), 425-434.

[SSD01] M. Stephens, N. J. Smith and P. Donnelly, A new statistical method for haplotype reconstruction from population data. *American Journal of Human Genetics* **68** (2001), 978-989.

[SSR95] E. Sim, L. A. Stanley, A. Risch and P. Thygesen, Xenogenetics in multifactorial disease susceptibility. *Trends in Genetics* **11** (1995), 509-512.

[Sta02] Statistics Finland, http://www.tilastokeskus.fi/tk/tp/tasku/taskus_vaesto.htm l#bkt [19.1.2002]

[STO01] P. Sevon, H. T. T. Toivonen and V. Ollikainen, TreeDT: Gene mapping by tree disequilibrium test. In *Proceeding of the*

Seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD-2001), 365 - 370, San Francisco, California, August 2001.

[StR96] T. Strachan and A. Read, Human Molecular Genetics. BIOS Scientific Publishers Limited, Oxford, 1996.

[SVV02] J. M. Satagopan, D. A. Verbel, E. S. Venkatraman, K. E. Offit and C. B. Begg, Two-stage designs for gene-disease association studies. *Biometrics* **58** (2002), 163-170.

[TeO92] J. D. Terwilliger and J. Ott, A haplotype-based 'haplotype relative risk' approach to detecting allelic associations. *Human Heredity* **42** (1992), 337-346.

[Ter95] J. Terwilliger, A powerful likelihood method for the analysis of linkage disequilibrium between trait loci and one or more polymorphic marker loci. *American Journal of Human Genetics* **56** (1995), 777-787.

[TeW98] J. Terwilliger and K. Weiss. Linkage disequilibrium mapping of complex disease: fantasy or reality? *Current Opinion in Biotechnology* **8** (1998), 578-594.

[ThN97] E. Thompson and J. Neel, Allelic disequilibrium and allele frequency distributions as a function of social and demographic history. *American Journal of Human Genetics* **60** (1997), 197-204.

[TOV00] H. T. T. Toivonen, P. Onkamo, K. Vasko, V. Ollikainen, P. Sevon, H. Mannila, M. Herr and J. Kere, Data mining applied to linkage disequilibrium mapping. *American Journal of Human Genetics* **67** (2000), 133-145.

[TOV00b] H. T. T. Toivonen, P. Onkamo, K. Vasko, V. Ollikainen, P. Sevon, H. Mannila and J. Kere, Gene Mapping by Haplotype Pattern Mining. In *Proceedings of IEEE International Symposium on Bio-Informatic and Biomedical Engineering*, 99-108, Arlington, VA, November 8-10, 2000.

[TSL97] J. D. Terwilliger, W. D. Shannon, G. M. Lathrop, J. P. Nolan, L. R. Goldin, G. A. Chase and D. E. Weeks, True and false positive peaks in genomewide scans: applications of length-biased sampling to linkage mapping. *American Journal of Human Genetics* **61** (1997), 430-438.

[TSO93] J. Terwilliger, M. Speer and J. Ott. Chromosome-based method for rapid computer simulation in human genetic linkage analyses. *Genetic Epidemiology* **10** (1993), 217-224.

[TZL98] J. Terwilliger, S. Zöllner, M. Laan and S. Pääbo, Mapping genes through the use of linkage disequilibrium generated by genetic drift: 'Drift mapping' in small populations with no demographic expansion. *Human Heredity* **48** (1998), 138-154.

[VaM97] M. A. van der Meulen and G. J. te Meerman, Association and haplotype sharing due to identity by descent, with an application to genetic mapping. In I.H. Pawlowitzki, J.H. Edwards and E.A. Thompson (eds.), Genetic Mapping of Disease Genes (1997), 115-135.

[VAM01] J. C. Venter, M. D. Adams, E. W. Myers, P. W. Li, R. J. Mural, G. G. Sutton, H. O. Smith, M. Yandell, C. A. Evans, R. A. Holt, et al., The Sequence of the Human Genome. *Science* **291** (2001), 1304-1351.

[WAA01] E. M. Wijsman, L. Almasy, C. I. Amos, I. Borecki, C. T. Falk, T. M. King, M. M. Martinez, D. Meyers, R. Neuman, J. M. Olson, S. Rich, M. A. Spence, D. C. Thomas, V. J. Vieland, J. S. Witte and J. W. MacCluer. Analysis of complex genetic traits: applications to asthma and simulated data. *Genetic Epidemiology* **21** (suppl. 1) (2001), S1-S853.

[Wei96] B. Weir, Genetic Data Analysis II. Sinauer Associates, Inc., Sunderland MA, 1996.

[WeL95] D. E. Weeks and M. Lathrop, Polygenic disease: methods for mapping complex disease traits. *Trends in Genetics* **11** (1995), 513-519.

[Woo55] B. Woolf, On estimating the relation between blood group and disease. *Annals of Human Genetics* **19** (1955), 251-253.

[WOL90] D. E. Weeks, J. Ott and G. M. Lathrop, SLINK: a general simulation program for linkage analysis. *American Journal of Human Genetics* **47** (1990), A204.

[WSL89] A. C. Warren, S. A. Slaugenhaupt, J. G. Lewis, A. Chakravarti, S. E. Antonarakis, A genetic linkage map of 17 markers on human chromosome 21. *Genomics* **4** (1989), 579-591.

[XiJ99] M. Xiong and L. Jin, Comparison of the power and accuracy of biallelic and microsatellite markers in population-based gene-mapping methods. *American Journal of Human Genetics* **64** (1999), 629-640.

[ZAH98] L. P. Zhao, C. Aragaki, L. Hsu and F. Quiaoit, Mapping of complex traits by single-nucleotide polymorphisms. *American Journal of Human Genetics* **63** (1998), 225-240.

Appendix A

Legend for Figure 4.4 (page 37) :

An example of inheritance of chromosomal segments in a pedigree of 9 individuals. Each of the four founder individuals (numbers 1, 2, 3, and 4) has two continuous (black and white) chromosomal segments, which are split by subsequent meioses. The paternally inherited chromosome of an individual is on the left and the maternally inherited on the right. The founder origin of each chromosomal segment is indicated next to each segment: for instance '2F' refers to the paternal chromosome of individual 2. At the bottom of the figure, an input file describing the structure of the pedigree is shown. For each individual, the input file contains family identifier, person identifier, father and mother identifiers, sex (1=male, 2=female), output status ('+'=output, '-'=do not output) and number of children (optional, used for memory management). In addition to the pedigree file, the user enters the genetic length of the region (in this case, 100 cM for both sexes). The output file in the example corresponds to one possible outcome of the process (for individuals 8 and 9), also illustrated graphically in the figure. In the output file, each row corresponds to a chromosomal segment and contains the person identifier, phase (0=paternally, 1=maternally inherited chromosome), segment endpoint, person identifier and phase (0 or 1) of the founder chromosome.

Legend for Figure 7.6 (page 106) :

Optimal power and false positive rates for Δ measure of association using haplotypes of four adjacent markers. Different sample sizes are denoted in the legend by A-B, where A stands for the number of mutation-carriers and B for the number of phenocopies in the affected sample. The corresponding marker maps are denoted by capital letters: (A) microsatellite map, 2 cM spacing, PIC=0.7 (B) microsatellite map, 3 cM spacing, PIC=0.7, (C) microsatellite map, 5 cM spacing, PIC=0.7, (D) SNP map, 1 cM spacing, allele frequencies 0.5/0.5, (E) SNP map, 2 cM spacing, allele frequencies 0.5/0.5, (F) SNP map, 1 cM spacing, allele frequencies 0.75/0.25, and (G) SNP map, 2 cM spacing, allele frequencies 0.75/0.25. The figures at the top correspond to populations BASIC (on the left) and IMM (on the right), and the figures at the bottom to populations DEATH (on the left) and STABLE (on the right).

Legends for the tables in this appendix:

Table 1 The statistic values for IBD sharing. The sample was collected from pseudocontrol chromosomes of random individuals (chromosome class 1). The column labeled 'Pop' refers to the population number (see Table 6.1 on page 79). The overall averages ('Avg') and standard deviations ('Std') of the sharing statistics are given, summarizing each one over the 8 locations and the 100 data sets.

Table 2 The statistic values for IBD sharing. This time, the sample was collected from pseudocontrol chromosomes of affected individuals (chromosome class 2).

Table 3 The statistic values for IBD sharing. The sample was collected from the chromosomes of affected individuals (chromosome class 3).

Pop	NUMCHR		AVGLEN		MAXLEN		AVGINH		EX5	
	Avg	Std	Avg	Std	Avg	Std	Avg	Std	Avg	Std
1	120.2	4.9	0.099	0.0037	0.421	0.059	3.3	0.14	295.4	10.2
2	127.4	5.0	0.100	0.0042	0.430	0.062	3.1	0.12	293.9	10.0
3	124.2	4.7	0.101	0.0042	0.427	0.060	3.2	0.12	297.0	9.7
4	130.3	5.7	0.103	0.0047	0.478	0.084	3.1	0.14	298.2	9.9
5	126.1	4.6	0.102	0.0039	0.433	0.075	3.2	0.12	297.4	8.6
6	73.1	3.1	0.102	0.0040	0.457	0.083	5.5	0.24	295.6	9.6
7	130.9	4.8	0.102	0.0043	0.444	0.073	3.1	0.11	296.6	10.2
8	201.0	6.8	0.099	0.0037	0.413	0.062	2.0	0.07	295.0	9.9
9	128.7	5.0	0.103	0.0040	0.450	0.070	3.1	0.12	300.8	9.7
10	145.0	6.5	0.135	0.0063	0.565	0.078	2.8	0.12	335.0	7.5
11	110.3	5.2	0.081	0.0033	0.341	0.043	3.6	0.17	261.6	12.4
12	96.5	4.3	0.067	0.0028	0.295	0.045	4.2	0.18	224.4	12.7
13	71.9	3.6	0.041	0.0017	0.173	0.025	5.6	0.27	117.6	10.4

Pop	EX10		EX15		EX20		EX25		EX30	
	Avg	Std	Avg	Std	Avg	Std	Avg	Std	Avg	Std
1	161.4	11.0	78.3	8.9	33.8	6.0	14.5	4.4	6.2	2.7
2	163.8	10.5	80.7	8.5	36.3	7.2	15.7	4.6	6.5	2.8
3	167.1	10.7	82.7	8.4	37.6	7.8	15.9	4.7	6.5	2.8
4	168.6	11.8	86.4	9.1	39.3	7.3	17.8	5.3	7.8	3.4
5	167.7	10.9	85.4	8.8	38.7	6.6	17.1	4.5	6.9	2.7
6	166.7	10.6	83.1	9.0	38.4	6.7	17.7	4.6	7.5	3.0
7	167.0	10.5	84.2	8.4	38.7	7.3	17.2	5.6	7.6	3.5
8	162.3	10.6	80.6	8.4	35.9	6.1	14.6	4.1	5.7	2.6
9	171.1	11.0	87.0	8.2	38.5	6.8	17.2	5.1	7.4	3.3
10	230.0	10.7	144.1	10.3	81.6	11.5	45.2	9.6	23.8	6.7
11	117.2	10.5	46.1	7.0	16.0	4.4	5.8	2.6	1.8	1.3
12	82.4	8.8	26.5	5.2	7.3	2.8	2.0	1.5	0.5	0.8
13	17.0	4.5	1.9	1.5	0.1	0.4	0.0	0.0	0.0	0.0

Pop	MAXINH		INH2		INH3		INH5		INH10	
	Avg	Std	Avg	Std	Avg	Std	Avg	Std	Avg	Std
1	15.3	2.7	84.5	4.7	58.9	3.9	28.7	2.9	4.8	1.7
2	13.6	2.2	88.1	4.6	60.4	3.8	27.9	2.8	3.7	1.5
3	14.9	3.1	87.0	4.3	60.2	3.5	28.3	2.9	3.9	1.5
4	12.4	2.0	91.4	4.7	62.4	3.8	27.4	3.0	2.7	1.5
5	14.4	3.2	89.4	4.1	61.8	3.8	27.6	3.0	3.5	1.5
6	21.7	4.0	60.1	3.3	49.4	3.2	33.3	2.8	12.5	1.9
7	14.2	2.6	89.4	4.4	60.2	3.9	26.5	3.0	3.5	1.4
8	8.4	1.6	102.5	4.9	50.5	3.9	11.5	2.6	0.2	0.5
9	12.5	2.3	90.0	4.2	62.0	3.9	28.1	2.9	2.8	1.5
10	12.0	2.2	94.9	4.6	60.6	3.9	23.8	3.1	2.0	1.2
11	17.0	3.8	80.8	4.7	58.3	4.1	30.0	2.9	6.0	1.9
12	16.3	2.8	75.1	4.1	57.9	3.9	33.6	2.9	7.4	2.0
13	23.8	5.2	60.1	3.5	49.7	3.3	33.1	2.7	12.1	1.9

Table 1: See page A–2.

Pop	NUMCHR		AVGLEN		MAXLEN		AVGINH		EX5	
	Avg	Std	Avg	Std	Avg	Std	Avg	Std	Avg	Std
1	125.3	4.9	0.098	0.0036	0.407	0.053	3.2	0.13	292.5	9.4
2	128.7	5.0	0.100	0.0039	0.418	0.066	3.1	0.12	295.7	9.7
3	126.8	5.3	0.101	0.0046	0.433	0.069	3.2	0.13	296.9	11.0
4	123.7	5.8	0.102	0.0040	0.433	0.065	3.2	0.15	295.3	10.2
5	126.5	5.3	0.103	0.0039	0.441	0.071	3.2	0.13	300.3	9.7
6	73.5	3.3	0.102	0.0040	0.442	0.071	5.5	0.24	296.3	9.9
7	125.1	5.0	0.102	0.0040	0.435	0.072	3.2	0.13	298.9	9.7
8	199.9	6.7	0.101	0.0038	0.429	0.058	2.0	0.07	295.2	8.8
9	123.4	6.2	0.104	0.0041	0.452	0.070	3.2	0.16	302.2	9.2
10	145.0	6.3	0.135	0.0058	0.573	0.090	2.8	0.12	332.0	8.6
11	107.2	5.0	0.080	0.0038	0.336	0.048	3.7	0.18	260.1	11.1
12	99.5	5.0	0.068	0.0026	0.290	0.040	4.0	0.20	228.1	11.6
13	66.7	4.2	0.041	0.0019	0.183	0.032	6.0	0.38	120.3	12.0

Pop	EX10		EX15		EX20		EX25		EX30	
	Avg	Std	Avg	Std	Avg	Std	Avg	Std	Avg	Std
1	160.4	10.3	77.9	8.0	33.7	6.0	14.0	3.9	5.5	2.4
2	164.4	10.7	81.0	8.5	36.2	6.4	14.7	4.2	5.9	2.7
3	166.6	12.5	83.0	10.0	36.1	6.7	15.3	4.7	6.3	2.9
4	167.9	11.9	85.5	8.8	38.7	6.4	17.0	4.7	7.5	3.1
5	170.9	11.5	87.2	8.9	39.4	6.9	17.8	4.7	7.3	3.0
6	168.7	10.5	84.8	8.9	37.7	7.1	16.6	4.9	7.1	3.1
7	167.3	11.4	83.3	9.0	37.5	7.1	16.4	4.6	6.7	2.9
8	167.1	10.8	83.1	8.6	36.8	6.6	15.7	4.6	6.4	2.9
9	172.2	11.3	87.7	8.9	40.4	7.2	18.9	5.3	7.7	3.1
10	228.7	10.1	141.9	9.8	80.3	10.5	44.7	8.7	24.1	6.8
11	115.6	11.0	45.2	7.9	15.1	4.7	5.2	2.6	1.6	1.4
12	83.3	8.3	26.2	5.0	7.2	2.8	2.1	1.5	0.5	0.7
13	18.6	4.7	2.3	1.6	0.3	0.5	0.0	0.2	0.0	0.1

Pop	MAXINH		INH2		INH3		INH5		INH10	
	Avg	Std	Avg	Std	Avg	Std	Avg	Std	Avg	Std
1	13.4	2.5	89.1	4.2	62.0	3.6	28.5	2.9	3.2	1.5
2	13.3	2.9	89.7	4.3	61.6	3.8	27.8	2.9	3.0	1.5
3	13.8	2.8	87.9	4.6	60.7	3.9	28.3	2.8	3.5	1.6
4	16.1	4.0	85.1	5.1	58.1	4.4	27.5	3.0	4.8	1.8
5	15.6	4.3	87.0	5.0	59.0	4.3	27.2	2.9	4.1	1.7
6	23.2	5.4	60.9	3.6	50.0	3.7	33.2	3.2	11.7	1.9
7	14.9	3.3	87.5	4.7	60.2	3.8	27.8	3.0	3.9	1.6
8	8.5	1.5	101.7	4.9	50.8	3.9	12.1	2.6	0.2	0.5
9	17.3	5.8	84.7	6.2	57.8	5.1	27.1	3.2	4.7	1.9
10	12.6	2.9	94.6	5.1	60.3	3.8	23.5	2.9	2.2	1.3
11	17.4	4.1	78.1	4.2	57.2	3.8	31.1	2.9	6.5	1.7
12	17.9	3.8	75.4	4.3	57.1	3.5	32.2	2.9	7.3	1.8
13	26.0	6.5	56.3	4.2	47.5	3.9	33.4	3.2	12.7	2.1

Table 2: See page A–2.

Pop	NUMCHR		AVGLEN		MAXLEN		AVGINH		EX5	
	Avg	Std	Avg	Std	Avg	Std	Avg	Std	Avg	Std
1	122.9	5.5	0.101	0.0039	0.430	0.071	3.3	0.15	295.2	9.5
2	124.7	5.5	0.100	0.0036	0.414	0.051	3.2	0.14	294.5	10.0
3	121.9	5.7	0.102	0.0044	0.421	0.059	3.3	0.16	299.0	10.2
4	122.5	6.6	0.102	0.0039	0.435	0.065	3.3	0.17	296.6	9.3
5	125.3	5.9	0.106	0.0046	0.448	0.076	3.2	0.15	304.8	9.6
6	77.1	3.4	0.102	0.0052	0.444	0.089	5.2	0.23	298.3	11.3
7	126.5	5.3	0.102	0.0044	0.427	0.061	3.2	0.13	296.7	11.3
8	199.7	7.0	0.100	0.0037	0.432	0.064	2.0	0.07	295.5	9.8
9	127.9	6.5	0.107	0.0039	0.452	0.059	3.1	0.16	303.1	10.0
10	145.4	6.0	0.130	0.0051	0.540	0.087	2.8	0.12	329.2	8.9
11	113.7	5.5	0.084	0.0035	0.352	0.056	3.5	0.17	268.3	10.4
12	94.4	4.5	0.068	0.0032	0.291	0.040	4.2	0.21	229.6	12.6
13	69.9	4.9	0.042	0.0019	0.183	0.025	5.7	0.41	126.9	11.6

Pop	EX10		EX15		EX20		EX25		EX30	
	Avg	Std	Avg	Std	Avg	Std	Avg	Std	Avg	Std
1	167.1	11.0	82.8	8.3	36.7	6.6	16.0	4.8	6.6	3.0
2	164.5	10.8	80.7	8.4	35.8	6.8	15.5	4.6	6.2	2.8
3	169.8	11.5	83.9	9.4	37.7	7.5	16.8	5.0	6.9	3.0
4	167.7	10.7	85.1	8.7	38.3	7.3	17.0	5.0	7.4	3.1
5	176.2	11.4	92.8	9.3	43.5	7.5	20.5	5.4	9.0	3.5
6	166.6	12.8	84.5	10.0	36.7	7.7	16.3	5.3	6.6	3.1
7	165.5	11.4	84.1	8.8	39.2	7.1	17.2	5.3	7.1	3.3
8	164.0	10.4	81.3	8.1	35.5	6.4	15.1	4.4	6.5	2.7
9	180.0	11.0	95.1	8.7	44.3	7.1	20.5	5.8	8.7	3.7
10	221.7	11.8	135.9	10.6	75.0	9.9	39.9	7.9	19.9	5.6
11	123.4	10.4	50.8	7.7	18.4	5.0	6.6	2.9	1.9	1.4
12	84.2	10.5	25.8	5.5	6.9	2.9	1.9	1.4	0.5	0.7
13	20.6	5.0	2.7	1.7	0.2	0.5	0.0	0.1	0.0	0.0

Pop	MAXINH		INH2		INH3		INH5		INH10	
	Avg	Std	Avg	Std	Avg	Std	Avg	Std	Avg	Std
1	15.0	5.5	86.0	4.5	59.9	3.8	28.8	2.9	4.0	1.6
2	15.5	6.0	87.0	4.6	59.9	3.9	27.9	3.0	3.9	1.5
3	15.1	5.2	85.1	4.9	59.4	4.1	28.9	3.0	4.3	1.7
4	16.1	5.5	84.7	5.2	58.8	4.3	28.3	3.0	4.6	1.8
5	18.2	7.0	86.2	5.2	58.5	4.4	26.6	3.1	4.5	1.7
6	25.9	10.8	63.1	4.2	51.0	4.4	32.5	3.7	10.5	1.9
7	15.9	6.6	87.3	4.7	59.5	3.9	27.1	3.0	4.0	1.6
8	9.0	2.7	100.4	5.1	50.3	3.9	12.4	2.8	0.3	0.6
9	17.1	6.9	87.5	6.5	59.0	5.3	26.2	3.1	4.0	2.1
10	13.8	5.2	95.6	5.2	60.8	4.3	22.8	3.0	2.1	1.3
11	17.5	6.0	83.1	4.5	59.7	3.9	29.8	3.1	4.9	1.6
12	18.9	5.8	73.3	4.3	56.3	3.8	32.6	2.7	8.2	1.9
13	25.8	7.3	58.4	4.7	48.1	4.1	32.9	2.8	12.4	2.1

Table 3: See page A–2.

TIETOJENKÄSITTELYTIETEEN LAITOS
PL 26 (Teollisuuskatu 23)
00014 Helsingin yliopisto

DEPARTMENT OF COMPUTER SCIENCE
P.O. Box 26 (Teollisuuskatu 23)
FIN-00014 University of Helsinki, FINLAND

JULKAISUSARJA **A**

SERIES OF PUBLICATIONS **A**

Reports may be ordered from: Department of Computer Science, Library (A 214), P.O. Box 26, FIN-00014 University of Helsinki, FINLAND.

A-1989-1 G. Grahne: The problem of incomplete information in relational databases. 156 + 3 pp. (Ph.D. thesis).

A-1989-2 H. Tirri (ed.): Interoperability of heterogeneous information systems: final report of the COST 11^{ter} project. 110 pp.

A-1989-3 J. Tarhio & M. Tienari (eds.): Computer Science at the University of Helsinki. 57 pp.

A-1989-4 T. Alanko, J. Keskinen, P. Kutvonen, M. Mutka, & M. Tienari: The AHTO project: software technology for open distributed processing. 53 + 3 pp.

A-1989-5 N. Holsti: Script editing for recovery and reversal in textual user interfaces. 126 pp. (Ph.D. thesis).

A-1989-6 K.E.E. Raatikainen: Modelling and analysis techniques for capacity planning. 162 + 52 pp. (Ph.D. thesis).

A-1990-1 K. Pohjonen & J. Tarhio (toim./eds.): Tietojenkäsittelyopin laitoksen tutkimusraportteja 1988–89 – Research reports at the Department of Computer Science 1988–89. 27 pp.

A-1990-2 J. Kuittinen, O. Nurmi, S. Sippu & E. Soisalon-Soininen: Efficient implementation of loops in bottom-up evaluation of logic queries. 14 pp.

A-1990-3 J. Tarhio & E. Ukkonen: Approximate Boyer-Moore string matching. 27 pp.

A-1990-4 E. Ukkonen & D. Wood: Approximate string matching with suffix automata. 14 pp.

A-1990-5 T. Kerola: Qsolver – a modular environment for solving queueing network models. 15 pp.

A-1990-6 Ker-I Ko, P. Orponen, U. Schöning & O. Watanabe: Instance complexity. 24 pp.

A-1991-1 J. Paakki: Paradigms for attribute-grammar-based language implementation. 71 + 146 pp. (Ph.D. thesis).

A-1991-2 O. Nurmi & E. Soisalon-Soininen: Uncoupling updating and rebalancing in chromatic binary search trees. 12 pp.

A-1991-3 T. Elomaa & J. Kivinen: Learning decision trees from noisy examples. 15 pp.

A-1991-4 P. Kilpeläinen & H. Mannila: Ordered and unordered tree inclusion. 22 pp.

A-1991-5 A. Valmari: Compositional state space generation. 30 pp.

A-1991-6 J. Tarhio & M. Tienari (eds.): Computer Science at the University of Helsinki 1991. 66 pp.

A-1991-7 P. Jokinen, J. Tarhio & E. Ukkonen: A comparison of approximate string matching algorithms. 23 pp.

A-1992-1 J. Kivinen: Problems in computational learning theory. 27 + 64 pp. (Ph.D. thesis).

A-1992-2 K. Pohjonen & J. Tarhio (toim./eds.): Tietojenkäsittelyopin laitoksen tutkimusraportteja 1990–91 – Research reports at the Department of Computer Science 1990–91. 35 pp.

A-1992-3 Th. Eiter, P. Kilpeläinen & H. Mannila: Recognizing renamable generalized propositional Horn formulas is NP-complete. 11 pp.

A-1992-4 A. Valmari: Alleviating state explosion during verification of behavioural equivalence. 57 pp.

A-1992-5 P. Floréen: Computational complexity problems in neural associative memories. 128 + 8 pp. (Ph.D. thesis).

A-1992-6 P. Kilpeläinen: Tree matching problems with applications to structured text databases. 110 pp. (Ph.D. thesis).

A-1993-1 E. Ukkonen: On-line construction of suffix-trees. 15 pp.

A-1993-2 Alois P. Heinz: Efficient implementation of a neural net α-β-evaluator. 13 pp.

A-1994-1 J. Eloranta: Minimal transition systems with respect to divergence preserving behavioural equivalences. 162 pp. (Ph.D. thesis).

A-1994-2 K. Pohjonen (toim./ed.): Tietojenkäsittelyopin laitoksen julkaisut 1992–93 – Publications from the Department of Computer Science 1992–93. 58 s./pp.

A-1994-3 T. Kujala & M. Tienari (eds.): Computer Science at the University of Helsinki 1993. 95 pp.

A-1994-4 P. Floréen & P. Orponen: Complexity issues in discrete Hopfield networks. 54 pp.

A-1995-1 P. Myllymäki: Mapping Bayesian networks to stochastic neural networks: a foundation for hybrid Bayesian-neural systems. 93 pp. (Ph.D. thesis).

A-1996-1 R. Kaivola: Equivalences, preorders and compositional verification for linear time temporal logic and concurrent systems. 185 pp. (Ph.D. thesis).

A-1996-2 T. Elomaa: Tools and techniques for decision tree learning. 140 pp. (Ph.D. thesis).

A-1996-3 J. Tarhio & M. Tienari (eds.): Computer Science at the University of Helsinki 1996. 89 pp.

A-1996-4 H. Ahonen: Generating grammars for structured documents using grammatical inference methods. 107 pp. (Ph.D. thesis).

A-1996-5 H. Toivonen: Discovery of frequent patterns in large data collections. 116 pp. (Ph.D. thesis).

A-1997-1 H. Tirri: Plausible prediction by Bayesian inference. 158 pp. (Ph.D. thesis).

A-1997-2 G. Lindén: Structured document transformations. 122 pp. (Ph.D. thesis).

A-1997-3 M. Nykänen: Querying string databases with modal logic. 150 pp. (Ph.D. thesis).

A-1997-4 E. Sutinen, J. Tarhio, S.-P. Lahtinen, A.-P. Tuovinen, E. Rautama & V. Meisalo: Eliot – an algorithm animation environment. 49 pp.

A-1998-1 G. Lindén & M. Tienari (eds.): Computer Science at the University of Helsinki 1998. 112 pp.

A-1998-2 L. Kutvonen: Trading services in open distributed environments. 231 + 6 pp. (Ph.D. thesis).

A-1998-3 E. Sutinen: Approximate pattern matching with the q-gram family. 116 pp. (Ph.D. thesis).

A-1999-1 M. Klemettinen: A knowledge discovery methodology for telecommunication network alarm databases. 137 pp. (Ph.D. thesis).

A-1999-2 J. Puustjärvi: Transactional workflows. 104 pp. (Ph.D. thesis).

A-1999-3 G. Lindén & E. Ukkonen (eds.): Department of Computer Science: annual report 1998. 55 pp.

A-1999-4 J. Kärkkäinen: Repetition-based text indexes. 106 pp. (Ph.D. thesis).

A-2000-1 P. Moen: Attribute, event eequence, and event type similarity notions for data mining. 190+9 pp. (Ph.D. thesis).

A-2000-2 B. Heikkinen: Generalization of document structures and document assembly. 179 pp. (Ph.D. thesis).

A-2000-3 P. Kähkipuro: Performance modeling framework for CORBA based distributed systems. 151+15 pp. (Ph.D. thesis).

A-2000-4 K. Lemström: String matching techniques for music retrieval. 56+56 pp. (Ph.D.Thesis).

A-2000-5 T. Karvi: Partially defined Lotos specifications and their refinement relations. 157 pp. (Ph.D.Thesis).

A-2001-1 J. Rousu: Efficient range partitioning in classification learning. 68+74 pp. (Ph.D. thesis)

A-2001-2 M. Salmenkivi: Computational methods for intensity models. 145 pp. (Ph.D. thesis)

A-2001-3 K. Fredriksson: Rotation invariant template matching. 138 pp. (Ph.D. thesis)

A-2002-1 A.-P. Tuovinen: Object-oriented engineering of visual languages. 185 pp. (Ph.D. thesis)

ISSN 1238-8645
ISBN 952-10-0614-5
Helsinki 2002
Helsinki University Printing House

Vesa Ollikainen: Simulation Techniques for Disease Gene Localization in Isolated Populations **A-2002-2**