
Preface

The ultimate goal of gene mapping is to identify the genes that play important roles in the inheritance of particular traits (phenotypes) and to explain the role of those genes in relation to one another and in relation to the environment. In practice this ambition is often reduced to identifying the genomic neighborhoods where one or a small number of the important genetic contributors to the phenotype are located.

Gene mapping takes place in many different organisms for many different reasons. In humans one is particularly interested in inherited or partly inherited diseases, and hopes that an identification of the responsible genes can lead in the relatively short run to better diagnostics and in the long run to strategies to alleviate the disease. In these cases the phenotype can be qualitative, whether an individual is affected with a particular disease, or it can be quantitative, say the level of a biomarker like cholesterol level, blood pressure, or body mass index, which is known or thought to be related to the disease. In plants or animals gene mapping can be of interest in its own right, or to produce more vigorous hybrid plants of agricultural value or farm animals that are more productive or more disease resistant. It can also involve model organisms, e.g., the plant *arabidopsis*, inbred strains of mice, or baker's yeast (*S. cerevisiae*), where one hopes to gain basic knowledge yielding insights that are broadly applicable.

The traits, or phenotypes, can be essentially any reproducible quality of an organism, and the goal of the mapping need not even be an actual gene. An example of considerable recent interest is a phenotype measuring the level of expression of some gene or genes under given experimental conditions, with the goal of discovering whether the expression of the gene is controlled by the immediate "upstream" region of the gene (*cis* control) or by some other genomic region, perhaps a master control region for a number of genes that must work in coordination (*trans* control).

There is variability in the expression of essentially any phenotype. There is also variability in the inheritance of genotypes, first because of Mendel's laws, but equally important for gene mapping because of recombination. The level of

variability is such that gene mapping necessarily has a statistical component, which is the subject of this book. At its simplest, gene mapping involves the correlation of the phenotype with the genotype of genetic markers, which are themselves hoped to be located close to, hence correlated with, the genes (or genomic regions) of interest.

Although gene mapping was practiced for a good part of the twentieth century, the subject has changed and grown substantially since the late 1980s. In the mid twentieth century the number of suitable markers was small, on the order of a few per genome (and the genomic location of these markers was often imprecise). As a consequence the principal impediment to gene mapping was the usually large genomic distance between gene and marker, which leads to small correlations between phenotype and marker genotype. The statistical model developed in human genetics to deal with this situation assumed that the mode of inheritance of a trait could be adequately modeled. The model almost invariably involved a single gene with the mode of inheritance, usually dominant or recessive, assumed to be known, and the penetrance, i.e., the conditional probability of expressing the trait given the genotype, also known. The unknown parameter of interest was the genetic distance from gene to marker, as measured by the recombination fraction, which then allowed one to test whether the trait was unlinked (recombination fraction = $1/2$) or linked (recombination fraction $< 1/2$) and to estimate the recombination fraction.

Since the explosion in the experimental techniques of molecular genetics in the late twentieth century, it has become possible to cover the genome with informative markers. As a consequence genes associated with many simple traits, which generate a large correlation between phenotype and markers that are close to the gene, have been mapped successfully. For complex traits, which may involve multiple genes, it is reasonable to assume that there is some marker close to the gene or genes influencing the trait, but the contribution of any particular gene may be small. This leads to small correlations between marker and phenotype, even if the marker is close to the gene; and this small correlation has now become the primary impediment to successful gene mapping.

The principal goal of this book is to explain the statistical principles of and problems arising in gene mapping. Particular emphasis is placed on the ideas that have arisen with the recent experimental developments leading to the availability of large numbers of molecular markers having known genomic locations and the desire to map progressively more complex traits. Indeed, the so-called "parametric" or "LOD" score method of human genetics, which was the established paradigm in human genetics from the time of the classical paper of N. Morton in 1955 [56] until quite recently and is still frequently used, is mentioned only briefly. (See Ott [57] for a thorough discussion.)

We have attempted to keep the formal statistical and computational requirements for reading the book as few as seems reasonable, with the hope that the book can be understood by a diverse audience. It is not, however, a handbook of methods, but a discussion of concepts, where we assume the

reader wants to play an active role, particularly in performing computational experiments and in thinking about the meaning of the results. Mathematical details are omitted when we did not think they gave added insight into the scientific issues. Since they can range from routine to difficult, we do not recommend that the reader expend effort to fill in the omitted details, unless he or she finds that intellectual activity rewarding for its own sake.

The book is organized globally as follows. The first three chapters deal with basic statistical, computational, and genetic concepts that are used later in the book. The next five are concerned with mapping quantitative traits using data from crosses of inbred strains. Chapters 9–13 involve primarily human genetics. Of those, Chaps. 9 and 11 discuss gene mapping based on data from pedigrees. This is conceptually similar to the first half of the book although necessarily substantially more complicated. Chapters 12 and 13 discuss association analysis, where relations between individuals in pedigrees are replaced by relations in populations. The discussion here is substantially less complete, and is to some extent limited to pointing out certain difficulties arising from complicated and uncertain population history. Chapter 10 involves admixture mapping, which has some features in common with the earlier chapters on gene mapping based on meiotic recombination and others related to population based association analysis.

A more detailed road map is as follows.

Chapter 1 reviews basic statistical concepts that are used throughout the book and explores these concepts computationally. It can be skipped or read quickly by someone familiar with basic statistics and computation in the language R.

In Chap. 2, we introduce our basic model relating the phenotype as dependent variable to genotype(s) as independent variable(s). It is a simple linear regression model having its origins in the classical paper of Fisher [30]. A principal attraction of the model is that straightforward variations can be developed to deal with quantitative or qualitative traits, which can be dominant or recessive, and which can involve multiple genes that interact with one another and/or with the environment. These variations are discussed at different places in the book.

Chapter 3 deals with some fundamental concepts of population genetics, and provides an opportunity to introduce some new programming techniques. It contains some difficult material, and except for recombination, which plays a central role throughout the book, most of the chapter is referenced only occasionally. The reader may wish to read Chap. 3 selectively, to get a rough idea of its contents and be prepared to refer to it occasionally.

For the experimental genetics of crosses of inbred lines, one can use the regression model and standard statistical methods to develop direct robust tests for linkage between phenotype and marker genotype. In Chap. 4 we discuss the simplest case of testing a single marker, first when that marker is itself a gene affecting the trait, and then when the marker is linked to a gene affecting the trait; and we see quantitatively the (deleterious) effect of

recombination between gene and marker on the power to determine if the marker is linked.

Since we will usually have no good idea where a gene affecting the trait is likely to be found, in Chap. 5 we introduce the notion of a genome scan, where we test a large number of markers distributed throughout the genome for linkage. This leads naturally to a problem of multiple comparisons that is solved both by computational methods and by theoretical results involving the maximum of a stochastic process. A systematic discussion of the power of a genome scan and of the related idea of confidence intervals for locating a linked gene as precisely as possible is given in Chap. 6.

Chapter 7 introduces in the simple context of experimental genetics the problem of missing information and one simple statistical idea to recapture that information. It turns out that in the context of that chapter, the solution is often more complicated than the problem warrants, but the problem appears again in a more complex form in Chaps. 9, 10, and 13, where missing information poses unavoidable difficulties that require progressively more complex algorithms to address.

Chapter 8 is concerned with more advanced problems in experimental genetics. Our goal here is to introduce the reader to these problems and point out which ones can in principle be dealt with by simple adaptations of methods developed up to that point and which ones pose more serious challenges.

Starting in Chap. 9, we discuss gene mapping in human genetics, where it is intrinsically more complicated, especially when there may be more than one gene and uncontrolled environmental conditions. Our discussion here is less complete than in the first eight chapters. It is essentially limited to pointing out how the theoretical framework of earlier chapters can be adapted to the more complex problems of human genetics and how the problem of missing information, which here moves to center stage, can be addressed. Chapters 9 and 11 are concerned with gene mapping based on inheritance within families. The concepts developed in Chaps. 4–8 are used, somewhat indirectly. Our presentation is designed to bring out the similarities while highlighting important differences. One important conclusion is that because of one's inability to perform breeding experiments in humans, family based methods for gene mapping are intrinsically less powerful than mapping in experimental genetics based on crosses of inbred lines.

Chapters 12 and 13 contain a brief introduction to gene mapping in populations, which is often called association analysis. It has the potential advantage over family based methods of substantially more power, providing one can successfully overcome some potential difficulties arising from the unknown population history. Our discussion is limited to describing a few simple models, the reasons they are attractive, and the potential pitfalls.

Chapter 10 is something of a bridge between the earlier chapters and the last two. While the methods discussed there are very similar to the methods of earlier chapters, and the issue of missing information is closely related to the same issue in Chap. 9, there are also complications of population history.

As indicated above, the classical parametric method of linkage analysis in human pedigrees is discussed only briefly. Also, in choosing to emphasize regression based methods, we have limited our discussion of likelihood methods, which can be very powerful when basic modeling assumptions are satisfied, to cases where they seemed to offer distinct advantages. The modeling assumptions, often in the form that phenotypes are normally distributed, can fail to hold, even approximately. When that happens, likelihood methods can be less robust than regression methods, although no statistical method should be regarded as so automatic that computer output is thought to speak for itself.

Finally, we would like to emphasize that the primary purpose of this book is didactic. The concepts and many related details have been published elsewhere by a large number of authors. We have provided some references to the scientific literature, largely for the purpose of introducing the reader to the substantial primary literature; but we have not tried to provide a complete scholarly bibliography.

For feedback in classes where preliminary versions of this book were used, we would like to thank students at The Hebrew University of Jerusalem, the Weizmann Institute, Stanford University, and the National University of Singapore. We also thank those universities, along with the Free University of Amsterdam, the Israel–U.S. Binational Science Foundation, the NIH, and the U.S. National Science Foundation for their support.

Stanford and Jerusalem,
October 2006

David O. Siegmund
Benjamin Yakir