Statistics for Biology and Health

Series Editors
M. Gail, K. Krickeberg, J. Samet, A. Tsiatis, W. Wong

Statistics for Biology and Health

Bacchieri/Cioppa: Fundamentals of Clinical Research
Borchers/Buckland/Zucchini: Estimating Animal Abundance: Closed Populations
Burzykowski/Molenberghs/Buyse: The Evaluation of Surrogate Endpoints
Cook/Lawless: The Statistical Analysis of Recurrent Events
Everitt/Rabe-Hesketh: Analyzing Medical Data Using S-PLUS
Ewens/Grant: Statistical Methods in Bioinformatics: An Introduction, 2nd ed.
Gentleman/Carey/Huber/Irizarry/Dudoit: Bioinformatics and Computational Biology Solutions Using R and Bioconductor
Hougaard: Analysis of Multivariate Survival Data
Keyfitz/Caswell: Applied Mathematical Demography, 3rd ed.
Klein/Moeschberger: Survival Analysis: Techniques for Censored and Truncated Data, 2nd ed.
Kleinbaum/Klein: Survival Analysis: A Self-Learning Text, 2nd ed.
Kleinbaum/Klein: Logistic Regression: A Self-Learning Text, 2nd ed.
Lange: Mathematical and Statistical Methods for Genetic Analysis, 2nd ed.
Manton/Singer/Suzman: Forecasting the Health of Elderly Populations
Martinussen/Scheike: Dynamic Regression Models for Survival Data
Moyé: Multiple Analyses in Clinical Trials: Fundamentals for Investigators
Nielsen: Statistical Methods in Molecular Evolution
O'Quigley: Proportional Hazards Regression
Parmigiani/Garrett/Irizarry/Zeger: The Analysis of Gene Expression Data: Methods and Software
Proschan/Lan Wittes: Statistical Monitoring of Clinical Trials: A Unified Approach
Siegmund/Yakir: The Statistics of Gene Mapping
Simon/Korn/McShane/Radmacher/Wright/Zhao: Design and Analysis of DNA Microarray Investigations
Sorensen/Gianola: Likelihood, Bayesian, and MCMC Methods in Quantitative Genetics
Stallard/Manton/Cohen: Forecasting Product Liability Claims: Epidemiology and Modeling in the Manville Asbestos Case
Sun: The Statistical Analysis of Interval-censored Failure Time Data
Therneau/Grambsch: Modeling Survival Data: Extending the Cox Model
Ting: Dose Finding in Drug Development
Vittinghoff/Glidden/Shiboski/McCulloch: Regression Methods in Biostatistics: Linear, Logistic, Survival, and Repeated Measures Models
Wu/Ma/Casella: Statistical Genetics of Quantitative Traits: Linkage, Maps, and QTL
Zhang/Singer: Recursive Partitioning in the Health Sciences
Zuur/Ieno/Smith: Analyzing Ecological Data

Rongling Wu
Chang-Xing Ma
George Casella

Statistical Genetics of Quantitative Traits

Linkage, Maps, and QTL

Springer

Rongling Wu
Department of Statistics
University of Florida
Gainesville, FL 32611
rwu@stat.ufl.edu

Chang-Xing Ma
Department of Biostatistics
State University of New York
at Buffalo
Buffalo, NY 14214
cxma@buffalo.edu

George Casella
Department of Statistics
University of Florida
Gainesville, FL 32611
casella@stat.ufl.edu

ISBN 978-0-387-20334-8 e-ISBN 978-0-387-68154-2

Library of Congress Control Number: 2006938665

Printed on acid-free paper.

9 8 7 6 5 4 3 2 1

springer.com

To my parents, wife, and son
RW

To my parents and Yuehua, David, and Eric
CM

To my family, and Lulu, too
GC

Preface

Most traits in nature and of importance to agriculture are quantitatively inherited. These traits are difficult to study due to the complex nature of their inheritance. However, recent developments of genomic technologies provide a revolutionary means for unraveling the secrets of genetic variation in quantitative traits. Genomic technologies allow the molecular characterization of polymorphic markers throughout the entire genome that are then used to identify and map the genes or quantitative trait loci (QTLs) underlying a quantitative trait based on linkage analysis.

Statistical analysis is a crucial tool for analyzing genome data, which are now becoming increasingly available for a variety of species, and for giving precise explanations regarding genetic variation in quantitative traits occurring among species, populations, families, and individuals. In 1989, Lander and Botstein published a hallmark methodological paper for interval mapping that enables geneticists to detect and estimate individual QTL that control the phenotype of a trait. Today, interval mapping is an important statistical tool for studying the genetics of quantitative traits at the molecular level, and has led to the discovery of thousands of QTLs responsible for a variety of traits in plants, animals, and humans. In a recent study published in *Science*, Li, Zhou, and Sang (2006, *311*, 1936–1939) were able to characterize the molecular basis of the reduction of grain shattering – a fundamental selection process for rice domestication – at the detected QTL by interval mapping. Among many other examples of the success of interval mapping are the positional cloning of QTLs responsible for fruit size and shape in tomato (Frary et al. 2000, *Science 289*, 85–88) and for branch, florescence, and grain architecture in maize (Doebley et al. 1997, *Nature 386*, 485–488; Gallavotti et al. 2004, *Nature 432*, 630–635; Wang et al. 2005, *Nature 436*, 714–719).

To make it suitable for various practical applications, interval mapping has been extensively modified and extended during the past 15 years. A host of useful statistical methods for QTL mapping have been produced through the collective efforts of statistical geneticists. However, these methods generally have various objectives and utilities and are sporadically distributed in a massive amount of literature. A single volume synthesizing statistical developments for genetic mapping may be helpful for many researchers, especially those with a keen interest in building a bridge between

genetics and statistics, to acquaint themselves with this expanding area as quickly as possible.

This book intends to provide geneticists with the tools needed to understand and model the genetic variation for quantitative traits based on genomic data collected in mapping research and equip statisticians with the uniqueness and ideas in relation to the exploration of genetic secrets using their computational skills. This book also intends to attract researchers toward multidisciplinary research and to introduce them to new paradigms in genomic science. In this book, the statistical and computational theories applied to genetic mapping are developed hand in hand and a number of examples displaying the implications of statistical genomics are introduced.

This book contains 14 chapters, broadly divided into three parts. Part 1, including Chapters 1 and 2, provides introductory genetics and statistics at the level appropriate for understanding general genetic concepts and statistical models for genetic mapping. Part 2, composed of Chapters 3–7, attempts to provide a thorough and comprehensive coverage of linkage analysis with molecular markers. Models and methods for linkage analysis and map construction are systematically introduced for different designs, such as the backcross/F_2 (Chapter 3), outbred crosses (Chapter 4), recombinant inbred lines (Chapter 5) and structured pedigrees (Chapter 7), and for special marker types including distorted and misclassified markers (Chapter 6) and dominant markers (Chapters 4 and 7). Part 3, composed of Chapters 8–14, covers statistical models and algorithms of QTL mapping. The topics include simple marker-phenotype association analyses (Chapter 8), the statistical structure of interval mapping (Chapter 9), regression- (Chapter 10) and maximum likelihood-based analysis of interval mapping (Chapter 11), threshold and confidence interval determination (Chapter 12), composite interval mapping using multiple markers as cofactors (Chapter 13), and interval mapping for outbred mapping populations (Chapter 14). In the Appendices, we provide general statistical theories directly related to the genetic mapping approaches introduced and R programs for some of the examples used in the book. A webpage (http://www.buffalo.edu/∼cxma/book/) was constructed for this book, which includes a complete list of programs and algorithms written in MatLab or R for all the examples.

Writing a book in such a rapidly developing and changing field is a pain but, more precisely speaking, full of excitement. In the summer of 1997, Wu delivered a series of lectures on statistical methods for QTL mapping to graduate students and faculty at Nanjing Forestry University, China. In the spring semester of 2002, Wu taught a statistical genetics course at the master's level at the University of Florida and then was joined for coteaching by Casella in the spring of 2003 and Ma in the spring of 2005. This course is now taught by Wu at the University of Florida and by Ma at the State University of New York at Buffalo on the regular basis. We all gave many lectures or short courses related to statistical genetics at other places and times. At each place and time, we were heavily impressed by the enthusiasm of students and other audiences to learn this fascinating area. All these encouraged us to write a book that can cover basic methods for statistical genetics research. The concepts, models and algorithms related to genetic mapping have been published in a variety of statistics and genetics journals by a large number of authors, but part of the material

contained in this book comes from our collaborative research program in the past five years. In particular, we apologize for those authors whose work was not mentioned in this book because of limited space.

During the writing of this book, many of our colleagues and friends both at the University of Florida and outside provided valuable help from different perspective. Wu is warmly grateful to his postdoctoral advisor, Dr. Zhao-Bang Zeng at North Carolina State University, for tremendous guidance and for leading him to the field of statistical genetics. Dr. Bruce Walsh at the University of Arizona provided insightful reviews of the book manuscript in different stages. Several anonymous reviewers gave constructive comments that significantly improve the presentation of the book. Students or postdocs who attended our lectures and classes or are working with us on statistical genetics in different places have provided many insightful suggestions to improve our presentation of the book. The following students or postdocs in our group, former or current, deserve special thanks: Yuehua Cui, Wei Hou, Hongying Li, Min Lin, Tian Liu, Fei Long, Xiang-Yang Lou, Qing Lu, Damaris Santana, Zhaojie Wang, Zuoheng Wang, Jiasheng Wu, Song Wu, Jie Yang, John Yap, Li Zhang, Wei Zhao, and Yun Zhu. The data used for examples in the book were kindly supplied by Dr. James Cheverud at Washington University (mouse), Dr. Junyi Gai at Nanjing Agricultural University (soybean), Rory Todhunter at Cornell University (dog), Drs. Stan Wullschleger and Tongming Yin at Oak Ridge National Laboratory (poplar), and Dr. Jun Zhu at Zhejiang University (rice).

We are grateful to the Department of Statistics and the Institute of Food and Agricultural Sciences at the University of Florida for writing this book and performing our research program. Finally, we are greatly indebted to our respective families for their continuous support of our research activities over the years. This work is partially supported by NSF grant 0540745.

Gainesville, FL — *Rongling Wu*
Buffalo, NY — *Chang-Xing Ma*
December 2006 — *George Casella*

Contents

1

Basic Genetics

1.1 Introduction

There have been enormous advances in the science of genetics. A huge amount of information regarding the precise molecular mechanisms of genetic transmission from parent to offspring is becoming increasingly available. In this chapter, we briefly review basic terminology and principles of genetics from Mendelian, population, quantitative and molecular perspectives at a level appropriate for understanding the research methods to be described in this book. Much of the description for classic Mendelian genetics is adapted from Bailey's (1961) book. To learn more about modern genetics, please look into the more general genetics textbooks that are listed at the end of this chapter.

1.2 Genes and Chromosomes

Genes are discrete units in which biological characteristics are inherited from parents to offspring. Genes are normally transmitted unchanged from generation to generation, and they usually occur in pairs. If a given pair consists of similar genes, the individual is said to be *homozygous* for the gene in question, while if the genes are dissimilar, the individual is said to be *heterozygous*. For example, if we have two alternative genes, say A and a, there are two kinds of homozygotes, namely AA and aa, and one kind of heterozygote, namely Aa. These alternative genes are called *alleles*. With a single pair of alleles, there are three different kinds of possible organisms represented by the three *genotypes* AA, Aa, and aa.

Genes are generally very numerous, and situated within the cell nucleus, where they lie in linear order along microscopic bodies called *chromosomes*. The chromosomes occur in similar, or *homologous*, pairs, where the number of pairs is constant for each species. For example, *Drosophila* has 4 pairs of chromosomes, pine has 12, the house mouse has 20, humans have 23, etc. The totality of these pairs constitutes the *genome* of a particular organism. One of the chromosome pairs in the genome

are the sex chromosomes (typically denoted by **X** and **Y**) that determine genetic sex. The other pairs are *autosomes* which guide the expression of most other traits.

Each gene pair has a certain place or *locus* on a particular chromosome. Since the chromosomes occur in pairs, the loci and the genes occupying them also occur in pairs. Therefore, it is the loci that have the fixed linear order, although a given locus may be occupied by any gene from the series of alleles (more than two alleles or *multialleles*) determining a particular trait. The most important purpose of a genome mapping project is to locate the genes affecting trait expressions on chromosomes.

1.3 Meiosis

When ordinary body cells divide and multiply, the cell nucleus undergoes a process of division called *mitosis*, which results in the two daughter cells, each having a full set of paired chromosomes exactly like the parent cell. But in the production of reproductive cells or *gametes* (ova and spermatozoa), we have a different mechanism, called *meiosis*. This ensures that only one chromosome from each homologous pair passes into each gamete. It follows that gametes also possess only one gene from each gene pair. The number of chromosomes in a gamete is referred to as the *haploid* number, in contrast to the full complement possessed by a fertilized egg, or *zygote*, which is *diploid*.

A diagram is drawn to illustrate the biological process of meiosis (Fig. 1.1). The chromosomes are already duplicated by the time they become visible at the start of the first meiotic division. Each pair of duplicates is joined at the *centromere*, a small particle at which two arms of the chromosome are connected. The duplicated pairs remain joined throughout the first anaphase. The paternal *homolog* (a duplicated pair) moves to one pole; the maternal homolog (another duplicated pair) moves to the other. The immediate products of the first meiotic division are two cells, each containing a diploid chromosome set. However, each homologous pair of chromosomes in one of these cells is a pair of maternally originated chromosomes or a pair of paternally originated chromosomes. The assortment between the two cells is random, with each resulting cell normally containing some chromosome pairs of maternal origin and others of paternal origin. In the second meiotic division, the number of chromosomes is halved and each of the two products of the first division produces identical daughter cells with half the usual number of chromosomes.

The significance of reduction division in meiosis is that it can maintain a diploid (double) chromosome set after fertilization, the fusion of a male gamete (sperm) with a female gamete (egg). A second essential characteristic of meiosis is that there is an interchange of genetic material between the two chromosomes of a homologous pair. Thus, the haploid gamete chromosome set contains a mixture of chromosomes, some derived from the father and some from the mother.

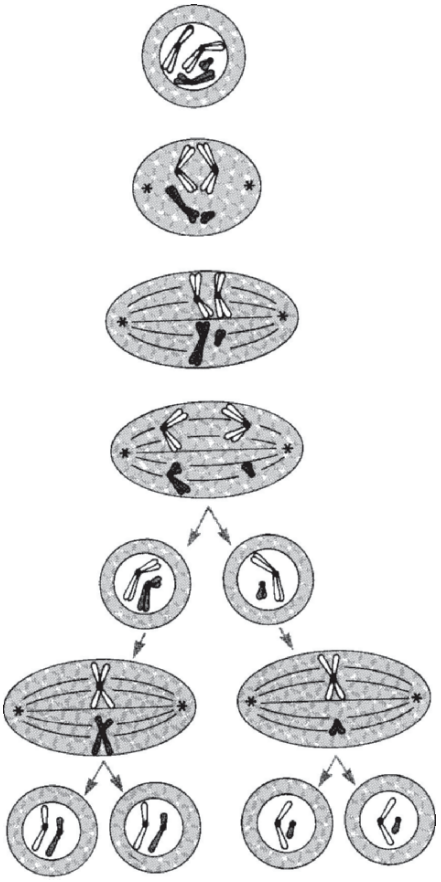

Fig. 1.1. Schematic diagram of meiosis in a hypothetical male who has one pair of identical autosomes (white) and one dissimilar XY pair (shaded). Adapted from Cavalli-Sforza and Bodmer (1971).

1.4 Mendel's Laws

1.4.1 Mendel's First Law

Genes are present in pairs in all cells of an adult organism, except for gametes. The gametes have only one gene from any given pair. Thus if an adult has genotype AA, all the gametes produced are of type A. But if the genotype is Aa, two types of gametes are possible, A and a, and these are normally produced in equal numbers. When fertilization occurs, a sperm carrying one gene from the male parent is united with an ovum carrying one gene from the female parent, thus making up a complete pair. The fertilized egg, or *zygote*, then develops to produce an organism in each body cell, of which one gene is derived from one parent and one from the other. The new individual produces its own reproductive cells, and so the process can continue.

The considerations above constitute Mendel's first law, the *Law of Segregation*. This states that characteristics are controlled by pairs of genes that segregate or separate during the formation of the reproductive cells, thus passing into different gametes. The pairs are restored when fertilization occurs, and this leads to the production of different types of offspring in certain definite proportions. In effect, segregation shuffles the genes and redeals them to the next generation. Characters themselves may also be said to show segregation, but the precise manner in which this happens depends on the nature of the genes involved and their dominant and recessive relationships.

Suppose we cross two individuals, represented by AA and aa. All gametes from the first will be A and all from the second will be a. Thus, all zygotes F_1 will be of the heterozygous type Aa. We now cross two individuals from the F_1 to form a new F_2 generation. Each F_1 heterozygous Aa produces two kinds of gametes, A and a, in equal numbers. At fertilization, there are four ways in which a zygote can be formed: one A gene from each parent; one a from each parent; A from the male and a from the female; or A from the female and a from the male. We therefore expect the three types of offspring AA, Aa, and aa in the ratios of 1:2:1 in the F_2 generation. But, if A is dominant, the first two classes will be phenotypically indistinguishable, giving the characters A and a in a 3:1 ratio.

If one of the heterozygous F_1 offspring is mated back to the homozygous parent, a *backcross* population is generated. The genotype of an individual in the backcross depends only on the heterozygous F_1 in which two kinds of gametes, A and a, are formed in equal numbers. Thus, the segregation ratio of the genotypes in the backcross follows a 1:1 ratio.

1.4.2 Mendel's Second Law

Mendel's second law says that when two or more pairs of genes segregate simultaneously, they do so independently. This is the *Law of Independent Assortment*. In some cases, this law is adequate, but it is subject to certain very important exceptions. These arise because of the phenomenon of linkage, a main topic of this book.

Suppose we have two pairs of genes represented by **A**, with two alleles A and a, and **B** with two alleles B and b. If we cross two individuals, one homozygous for both A and B and the other homozygous for both a and b (i.e., the mating $AABB \times aabb$), it is obvious that all offspring will be $AaBb$. This is because the first parent must produce gametes that are all AB, and the second parent must produce gametes which are all ab. We now consider the intercross $AaBb \times AaBb$. If the segregation is to be independent then each of these individuals will produce four kinds of gametes, namely AB, Ab, aB and ab, in equal numbers. Combining the four alternative types of gametes from one parent with the four alternatives from the other leads to 16 combinations, which are not, however, all different. The various possibilities are most easily presented as shown in the diagram of Fig. 1.2. It will be seen from the diagram that there are in fact nine distinct genotypes, $AABB$ (1), $AABb$ (2), $AAbb$ (1), $AaBB$ (2), $AaBb$ (4), $Aabb$ (2), $aaBB$ (1), $aaBb$ (2), and $aabb$ (1), where the number given in parentheses is the forming number of each genotype. But if each gene pair exhibits

a dominant/recessive relationship, there will be only four separate phenotypic classes, AB, Ab, aB, and ab occurring in the ratio 9:3:3:1.

Gametes	*AB*	*Ab*	*aB*	*ab*
AB	*AABB* (*AB*)	*AABb* (*AB*)	*AaBB* (*AB*)	*AaBb* (*AB*)
Ab	*AABb* (*AB*)	*AAbb* (*Ab*)	*AaBb* (*AB*)	*Aabb* (*Ab*)
aB	*AaBB* (*AB*)	*AaBb* (*AB*)	*aaBB* (*aB*)	*aaBb* (*aB*)
ab	*AaBb* (*AB*)	*Aabb* (*Ab*)	*aaBb* (*aB*)	*aabb* (*ab*)

Gametes

Fig. 1.2. Gene segregation of an intercross, $AaBb \times AaBb$, involving two gene pairs. When each pair exhibits dominance, the resultant phenotypes are given in brackets. The degree of dominance is roughly described by different darknesses of the cells.

1.5 Linkage and Mapping

Mendel's second law applies to genes whose loci lie on different chromosomes. Genes whose loci lie on the same chromosome will tend to remain together. Loci on the same chromosome are said to be *syntenic*, and those on different chromosomes are said to be *nonsyntenic*. The extent to which syntenic loci remain together depends on their closeness. We are thus led to consider the phenomenon of *linkage*.

In order to see what essentially is involved in linkage, let us consider the formation of gametes by a heterozygote $AaBb$. If the loci for the gene pairs A, a and B, b lie on the same kind of chromosome, we can specify more exactly the composition of the homologous pair of chromosomes. Thus, one chromosome may contain A and B, the other a and b; i.e.,

$$\begin{matrix} A & \vdots & a \\ B & \vdots & b \end{matrix} , \tag{1.1}$$

where the two vertical lines stand for the two homologous chromosomes. Or, alternatively, A and b may lie on one chromosome, while the other contains a and B; i.e.,

$$\begin{matrix} A & & a \\ b & & B \end{matrix} . \tag{1.2}$$

Definition 1.1. [Some Basic Terms] For alleles A and B, the arrangement displayed in diagram (1.1) is termed *coupling* and is written AB/ab; the arrangement in diagram (1.2) is called *repulsion* and is indicated by Ab/aB. The relative arrangement of *nonalleles* (i.e., A vs. B, A vs. b, a vs. B, or a vs. b) at different loci along a chromosome is called the *linkage phase*.

At an early stage of meiosis, the two chromosomes 1 and 2 lie side by side with corresponding loci aligned. If the parental genotype is AB/ab, we can represent the alignment as in Fig. 1.3A. Each of the paired chromosomes is then duplicated to form two sister strands (*chromatids*) connected to each other at a region called the *centromere*. The homologous chromosomes form pairs, so that each resulting complex consists of four chromatids known as a *tetrad* (Fig. 1.3B). At this stage, the non-sister chromatids adhere to each other in a semi-random fashion at regions called *chiasmata*. Each chiasma represents a point where *crossing over* between two non-sister chromatids can occur (Fig. 1.3C). Chiasmata do not occur entirely at random, as they are more likely farther away from the centromere, and it is unusual to find two chiasmata in very close proximity to each other.

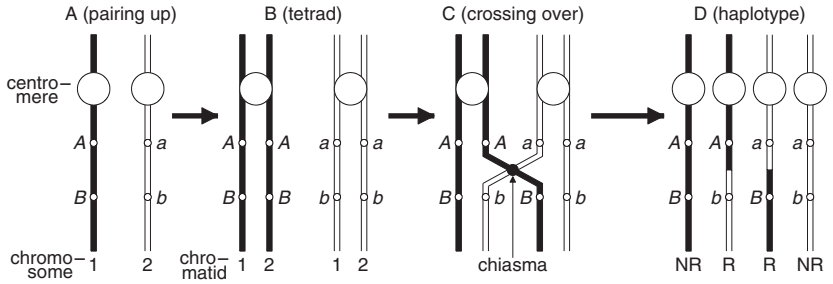

Fig. 1.3. Diagram for crossing–over between linked loci **A** and **B**.

Each gamete receives one chromatid from a tetrad to make up the haploid complement (Fig. 1.3D). Since it is possible that more than one crossover occurs on the chromosomes, some chromosomes in the haploid complement consist of a number of segments from the two parental chromosomes. The number of segments is determined by the number of crossovers that occurred in the formation of the chromatid that became the chromosome. If no crossovers occur, then the chromosome will be a replicate of an entire parental chromosome. If one crossover occurs between two loci $\mathcal{A}$ and $\mathcal{B}$, then the chromosome will consist of two segments, one from each parental chromosome. In the former case, the resultant gametes must be AB or ab, just like the *parental* chromosomes. In the latter case, where there is one point of exchange, we have the new combinations Ab and aB, called *recombinant* types. In general, if

there are an even number of points of exchange between the two loci, the final result will be indistinguishable from AB or ab. But if there are an odd number of points of exchange, the result will be like Ab or aB.

The existence of linkage means that there will be more gametes like AB and ab, and fewer like Ab and aB. Let us suppose that the proportion of recombinant gametes is r, which we call the *recombination fraction*, and that the proportion of parental type is $1 - r$. The recombination fraction can be estimated on the basis of the expected number of recombinants in a segregating progeny (see Chapter 3). In general, we should not expect to find recombination fractions greater than one-half, though in certain unusual circumstances there may be a tendency for chromosomes inherited from one parent or from particular stocks to associate nonrandomly.

From the definition of the recombination fraction, it follows that the special case $r = 1/2$ is equivalent to independent segregation or no linkage. Actually, if two loci on one chromosome are a long way apart, odd and even numbers of points of exchange will be about equally frequent (i.e., 50 percent each), so this case will not be immediately distinguishable from the case where the loci are on different chromosomes. Alternatively, if two loci are close together, the frequency of points of exchange will be low, and the corresponding recombination fraction will be small. To some extent, we can use the latter as a measure of the distance between any two loci.

A better scale of measurement is that afforded by the density of points of exchange.

Definition 1.2. [Map Distance] The *map distance* between any two loci is the average number of points of exchange occurring in the segment.

The map distance is a quantity that is automatically additive. There is a very simple relationship between the recombination fraction and the map distance for a pair of loci in the simplest case of no interference. Such a relationship is called a *map function* and will be discussed in Section 3.10. When the recombination fractions between pairs of loci on a single chromosome have been determined from an appropriate linkage experiment, it is a simple matter to transform them into map distances and hence construct a chromosome map. Since there is no reason to suppose that chromosomes are homogeneous along their lengths with regard to the frequency of crossing−over, we cannot assume that there is necessarily a very close correspondence between genetic map distance and the actual physical distance between the corresponding genes.

When many genes are considered, an issue arises about their linear arrangement within each chromosome. The loci of any organism fall into linkage groups, where any locus in one group is unlinked to any locus in a different group. Within any group, however, the loci can be arranged in a linear order. For sufficiently close loci, the recombination fraction between any pair may, in an elementary analysis, be used as a direct measure of the distance between the loci. To retain additivity at greater separations, we must work in terms of the average number of crossovers rather than the recombination fraction (which only measures the frequency of an odd number of crossovers). We thus need to know how the recombination fractions observed between many pairs of loci lying on a single chromosome can be fitted into a unifying picture

based on the notion of a chromosome *map*. This will critically rely upon the development of theoretical models and statistical algorithms for constructing genetic linkage maps, which is one of the major themes of this book.

1.6 Interference

In the simplest case, we assume that the points of exchange occur at random, so that the pattern of crossing−over in any segment of a chromosome is independent of the pattern in any other segment. In practice, however, nonrandomness is common and was named *interference* by H. J. Muller (1916). When, as usual, this is positive, the occurrence of a point of exchange tends to inhibit the formation of other such points in its neighborhood. Various models are available for describing the phenomenon of interference, and some of these entail the occurrence of recombination fractions greater than one-half in sufficiently long chromosomes.

As mentioned earlier, each chromosome splits longitudinally into a pair of identical daughter−chromosomes (chromatids) during the relevant part of a meiotic division. The two chromatids are initially held together by the centromere (Fig. 1.3B). Crossing−over always occurs between chromatids from different chromosomes of a homologous pair, as shown in Fig. 1.3C. Thus, the phenomenon of crossing−over actually involves all four chromatids, or strands, of any pair of homologous chromosomes. A pair of homologous chromosomes united by crossing−over is often called *bivalent*.

We may envision the occurrence of several points of exchange or chiasma, each of which now entails the X-like arrangement of chromatids shown in Fig. 1.3C. We can distinguish between two kinds of interference.

Definition 1.3. [Kinds of Interference] One type of interference is *chiasma interference*, in which the occurrence of one chiasma influences the chance of another occurring in its neighborhood, and another is *chromatid interference*, which is a nonrandom relationship between the pair of strands involved in one chiasma and the pair involved in the next chiasma.

Chiasma interference is common, and some distributions have been observed in which the variance of interference was as low as a quarter of its mean. Chromatid interference, on the other hand, is much more difficult to detect, and evidence for its existence is more scant. It has been proven that chiasma interference alone is incapable of causing recombination fractions of more than 50 percent (Mather 1938).

1.7 Quantitative Genetics

1.7.1 Population Properties of Genes

Mendelian segregation leads to simple and predictable segregation ratios in the offspring of specific mating types but only applies to a progeny population derived from

two parents of known genotype. However, different mating types can occur simultaneously to generate the offspring in a natural or experimental population in which the ratios of the different genotypes are weighted averages of the segregation ratios of all the possible mating types, the weights being the relative frequencies of the different mating types. The population properties of genes can be described by the allele frequencies, genotype frequencies, and Hardy-Weinberg law.

Consider a gene with two alleles, A and a, with respective frequencies p_1 and p_0, in a population. Let P_2, P_1, and P_0 be the population frequencies of three genotypes, AA, Aa and aa, respectively. When the mating type frequencies arise from random mating, the ratios of the different genotypes follow a mathematical model established independently by the English mathematician Hardy (1908) and the German physician Weinberg (1908). This well-known model, today called the Hardy-Weinberg Law, states that, if individuals in the population mated with each other at random, these frequencies would satisfy the relationship

$$P_1^2 = 4P_2P_0, \tag{1.3}$$

and each of these frequencies is kept unchanged from generation to generation. The population that follows equation (1.3) is said to be at Hardy-Weinberg equilibrium, in which the genotype frequencies can be expressed as $P_2 = p_1^2$, $P_1 = 2p_1p_0$, and $P_0 = p_0^2$, respectively. Approaches exist to test whether or not a population is at Hardy-Weinberg equilibrium (Falconer and Mackay 1996; Lynch and Walsh 1998).

1.7.2 A General Quantitative Genetic Model

A gene that is segregating in a population may affect the phenotype of a trait. For a complex or quantitatively inherited trait, the genes that determine it may be numerous and their relationships with the environment may be complicated. The study of the genetic basis of a quantitative trait is the theme of quantitative genetics.

Consider a quantitative trait with phenotypic value P, which is determined by the genetic (G) and environmental factors (E) and their interaction (G × E), expressed as

$$\mathrm{P} = \mathrm{G} + \mathrm{E} + \mathrm{G} \times \mathrm{E}. \tag{1.4}$$

Assuming that all terms in equation (1.4) are independent of one another, we partition the phenotypic variance of the trait into the corresponding genetic, environmental, and genotype × environment interaction variance components:

$$V_\mathrm{P} = V_\mathrm{G} + V_\mathrm{E} + V_{\mathrm{G}\times\mathrm{E}}. \tag{1.5}$$

In statistics, the variance is generally symbolized by V or σ^2. The genetic variance, V_G or σ_G^2, is due to the effects of all genes that determine the trait. Consider a gene with genotypes AA, Aa, and aa whose genotypic values and frequencies in a population at Hardy-Weinberg equilibrium are expressed as follows:

Genotype	Genotypic Value	Frequency
AA	$\mu_2 = \mu + a$	$P_2 = p_1^2$
Aa	$\mu_1 = \mu + d$	$P_1 = 2p_1p_0$
aa	$\mu_0 = \mu - a$	$P_0 = p_0^2$

The three different genotypes are symbolized by j ($j = 2$ for AA, 1 for Aa, and 0 for aa). Genotypic values are composed of the overall mean of the trait (μ), the *additive effect* (a) of the gene due to the substitution of alleles from A to a, or the *dominance effect* (d) due to the interaction effect of different alleles A and a at the gene. If there is no dominance, $d = 0$; if allele A is dominant over a, d is positive; and if allele a is dominant over A, d is negative. Dominance is complete if d is equal to $+a$ or $-a$, and there is overdominance if d is greater than $+a$ or less than $-a$. The degree of dominance is described by the ratio d/a.

The population mean of the three genotypes with different frequencies is calculated as

$$\bar{\mu} = \sum_{j=0}^{2} P_j \mu_j = (p_1 - p_0)a + 2p_1p_0d,$$

and we have the genetic variance for this gene,

$$\begin{aligned}
\sigma_g^2 &= \sum_{j=0}^{2} P_j \left(\mu_j - \bar{\mu}\right)^2 \\
&= 2p_1p_0[a + (p_1 - p_0)d]^2 + 4p_1^2p_0^2d^2 \\
&= 2p_1p_0\alpha^2 + 4p_1^2p_0^2d^2 \\
&\stackrel{def}{=} \sigma_a^2 + \sigma_d^2,
\end{aligned}$$

where $\alpha = a + (p_1 - p_0)d$ is the *average effect* due to the substitution of alleles from A to a (Falconer and Mackay 1996). The first term of the genetic variance, σ_A^2, is the *additive genetic variance* component, and the second term, σ_D^2, is the *dominance genetic variance* component. These two expressions can be readily extended to include the effects of all underlying genes for a trait. If gene interactions are ignored, the variances contributed by all the genes are expressed as $\sigma_G^2 = \sum \sigma_g^2$, $\sigma_A^2 = \sum \sigma_a^2$, and $\sigma_D^2 = \sum \sigma_d^2$.

1.7.3 Genetic Models for the Backcross and F_2 Design

The partitioning of the genetic variance can be made for different genetic settings. Consider two parental populations, P_1 and P_2, fixed with favorable alleles $A_1, ..., A_m$ and unfavorable alleles $a_1, ..., a_m$, respectively, for all m loci. The two parents are crossed to generate an F_1. The F_1 is backcrossed to one of the parents to form a backcross or self-crossed to form an F_2.

Let a_k and d_k be the additive and dominance effects of gene k, respectively, and r_{kl} be the recombination fraction between any two genes k and l. Consider a pair of genes, $\mathbf{A}_k$ and $\mathbf{A}_l$, whose genotypic values (upper) and frequencies (lower) in the F_2 population are expressed as

(1.6)

	A_lA_l	A_la_l	a_la_l
A_kA_k	$\mu + a_k + a_l$	$\mu + a_k + d_l$	$\mu + a_k - a_l$
	$\frac{1}{4}(1 - r_{kl})^2$	$\frac{1}{2}r_{kl}(1 - r_{kl})$	$\frac{1}{4}r_{kl}^2$
A_ka_k	$\mu + d_k + a_l$	$\mu + d_1 + d_2$	$\mu + d_k - a_l$
	$\frac{1}{2}r_{kl}(1 - r_{kl})$	$\frac{1}{2}[r_{kl}^2 + (1 - r_{kl})^2]$	$\frac{1}{2}r_{kl}(1 - r_{kl})^2$
a_ka_k	$\mu - a_k + a_l$	$\mu - a_k + d_l$	$\mu - a_k - a_l$
	$\frac{1}{4}r_{kl}^2$	$\frac{1}{2}r_{kl}(1 - r_{kl})$	$\frac{1}{4}(1 - r_{kl})^2$

where the genotypic values are composed of the additive and dominance effects at the two genes because gene interactions are ignored, and the derivation of the genotype frequencies in the F_2, expressed in terms of the recombination fraction between two genes, needs knowledge of linkage analysis, described in Section 3.5. From display 1.6, we can derive the genetic variance of the trait as

$$\begin{aligned}\sigma_G^2 = {} & \frac{1}{2}\sum_{k=1}^{m} a_k^2 + \frac{1}{4}\sum_{k=1}^{m} d_k^2 \\ & + \frac{1}{2}\sum_{k=1}^{m}\sum_{l=1,k\neq l}^{m}(1 - 2r_{kl})a_k a_l + \frac{1}{4}\sum_{k=1}^{m}\sum_{l=1,k\neq l}^{m}(1 - 2r_{kl})^2 d_k d_l.\end{aligned} \tag{1.7}$$

The first term on the right side of equation (1.7) for the F_2 is the additive variance within loci, the second is the dominance variance within loci, the third is the additive covariance between different loci, and the fourth is the dominance covariance between different loci.

For the backcross, in which the dominance effect cannot be defined due to inadequate degrees of freedom, we can derive a similar but simpler genetic variance, expressed as

$$\sigma_G^2 = \frac{1}{4}\sum_{k=1}^{m} a_k^2 + \frac{1}{4}\sum_{k\neq l}^{m}(1 - 2r_{kl})a_k a_l. \tag{1.8}$$

From equation (1.8), the genetic variance in a backcross consists of the additive genetic variance and additive covariance between different loci.

1.7.4 Epistatic Model

Genes may affect quantitative traits in an interactive way. The effect due to gene interaction was coined as *epistasis* by W. Bateson (1902). From a physiological perspective, epistasis describes the dependence of gene effects at one locus upon those at the other locus. Fisher (1918) first partitioned the genetic variance into additive, dominance, and epistatic components using the least squares principle. Cockerham (1954) further partitioned the two-gene epistatic variance into the additive × additive, additive × dominance, dominance × additive, and dominance × dominance interaction components. There are many approaches for specifying epistasis, but we will model epistasis using Mather and Jinks' (1982) approach.

Consider two genes, one denoted by **A**, with three genotypes, AA, Aa, and aa, and the second denoted by **B**, with three genotypes, BB, Bb, and bb. These two genes form nine two-locus genotypes, whose genotypic values, denoted by $\mu_{j_1 j_2}$, can be partitioned into different components

$$
\begin{aligned}
\mu_{j_1 j_2} = \quad & \mu && \text{overall mean} \\
& + (j_1 - 1)a_1 + (j_2 - 1)a_2 && \text{additive effects} \\
& + j_1(2 - j_1)d_1 + j_2(2 - j_2)d_2 && \text{dominance effects} \\
& + (j_1 - 1)(j_2 - 1)i_{aa} && \text{additive} \times \text{additive effect} \\
& + (j_1 - 1)j_2(2 - j_2)i_{ad} && \text{additive} \times \text{dominance effect} \\
& + j_1(2 - j_1)(j_2 - 1)i_{da} && \text{dominance} \times \text{additive effect} \\
& + j_1(2 - j_1)j_2(2 - j_2)i_{dd} && \text{dominance} \times \text{dominance effect},
\end{aligned}
\tag{1.9}
$$

where

$$
j_1, j_2 = \begin{cases} 2 & \text{for } AA \text{ or } BB \\ 1 & \text{for } Aa \text{ or } Bb \\ 0 & \text{for } aa \text{ or } bb \end{cases} .
$$

The second line of equation (1.9) is the additive effects of single genes, the third line is the dominance effects of single genes, and the fourth, fifth, sixth, and seventh lines are the epistatic effects between the two genes, additive × additive (i_{aa}), additive × dominance (i_{ad}), dominance × additive (i_{da}), and dominance × dominance (i_{dd}), respectively.

For the two genes that are cosegregating with the recombination fraction of r in an F_2 population, the genotypic values and frequencies are expressed in Table 1.1. Note that the genotype frequencies are calculated in terms of r. Based on Table 1.1, the genetic variance of a trait can be derived.

1.7.5 Heritability and Its Estimation

According to equation (1.5), the total phenotypic variance of a quantitative trait is decomposed into its genetic, environment and genotype × environment interaction

Table 1.1. Genotypic values (upper) and frequencies (lower) of the nine genotypes at two genes, **A** and **B**.

	BB	Bb	bb
AA	$\mu + a_1 + a_2 + i_{aa}$	$\mu + a_1 + d_2 + i_{ad}$	$\mu + a_1 - a_2 - i_{aa}$
	$\frac{1}{4}(1-r)^2$	$\frac{1}{2}r(1-r)$	$\frac{1}{4}r^2$
Aa	$\mu + d_1 + a_2 + i_{da}$	$\mu + d_1 + d_2 + i_{dd}$	$\mu + d_1 - a_2 - i_{da}$
	$\frac{1}{2}r(1-r)$	$\frac{1}{2}[r^2 + (1-r)^2]$	$\frac{1}{4}r(1-r)$
aa	$\mu - a_1 + a_2 - i_{aa}$	$\mu - a_1 + d_2 - i_{ad}$	$\mu - a_1 - a_2 + i_{aa}$
	$\frac{1}{4}r^2$	$\frac{1}{2}r(1-r)$	$\frac{1}{4}(1-r)^2$

variance components. The ratio of the genetic variance over the phenotypic variance is defined as *broad-sense heritability*, i.e.,

$$H^2 = \frac{\mathrm{V_G}}{\mathrm{V_G} + \mathrm{V_E} + \mathrm{V_{G\times E}}}. \tag{1.10}$$

As shown above, the genetic effect or variance can be partitioned into additive (A) and nonadditive (NA) effects or variances. Thus, we have

$$\begin{aligned} \mathrm{P} &= \mathrm{G} + \mathrm{E} + \mathrm{G}\times\mathrm{E} \\ &= \mathrm{A} + \mathrm{NA} + \mathrm{E} + \mathrm{A}\times\mathrm{E} + \mathrm{NA}\times\mathrm{E}, \end{aligned}$$

and

$$\begin{aligned} V_\mathrm{P} &= V_\mathrm{G} + V_\mathrm{E} + V_{\mathrm{G}\times\mathrm{E}} \\ &= V_\mathrm{A} + V_\mathrm{NA} + V_\mathrm{E} + V_{\mathrm{A}\times\mathrm{E}} + V_{\mathrm{NA}\times\mathrm{E}}, \end{aligned}$$

if all the effects terms are independent of each other.

The nonadditive effect or variance is the summation of dominance and epistatic effect or variance. Because the additive effect can be inherited from the parents to offspring whereas the nonadditive effect cannot, we use the ratio of the additive variance over the total phenotypic variance, define as the *narrow-sense heritability*, i.e.,

$$h^2 = \frac{\mathrm{V_A}}{\mathrm{V_A} + \mathrm{V_{NA}} + \mathrm{V_E} + \mathrm{V_{A\times E}} + \mathrm{V_{NA\times E}}}, \tag{1.11}$$

to quantify the degree with which the phenotypic value of a quantitative trait is unchanged from one generation to next. The two heritability parameters (1.10) and (1.11) are traditionally used to describe the degree of overall genetic control for a trait, including the contributions of all the underlying genes (Lynch and Walsh 1998). These

two parameters are now commonly used to describe the contributions of individual genes if these genes can be detected by an approach like genetic mapping, described in Chapters 8–14.

In practice, genetic variances can be estimated on the basis of a quantitative genetic theory founded by Cockerham (1954, 1963). According to this theory, a set of parents is crossed to generate multiple crosses in a mating design. The progeny from the mating design is then grown in a particular experimental design, from which the phenotypic data collected are analyzed by statistical approaches, such as analysis of variance, to obtain various experimental variances. Based on the resemblance between relatives, the estimated experimental variances are used to estimate the additive and dominance genetic variances and, therefore, the broad- and narrow-sense heritabilities.

Comparable to Cockerham's models, Mather and Jinks (1982) proposed a different approach based on generation differences to estimate genetic effect or variance components. Consider study material composed of three generations, inbred parents P_1 and P_2, the non-segregating F_1 and the segregating F_2, which are grown under the same condition. The phenotypic variance of a trait for the two pure parent lines (V_{P_1} and V_{P_2}) and F_1 progeny (V_{F_1}) is purely due to environmental factors, whereas the phenotypic variance of the same trait in the F_2 (V_{F_2}) includes a sum of genetic, environmental and genotype × environmental variance. Thus, the genetic variance of the trait can be estimated by

$$V_G = V_{F_2} - V_{F_1}, \tag{1.12}$$

or

$$V_G = V_{F_2} - \frac{1}{4}\left(V_{P_1} + V_{P_2} + 2V_{F_1}\right). \tag{1.13}$$

The estimates of individual genetic variance components can be obtained by the inclusion of more generations (Mather and Jinks 1982).

1.7.6 Genetic Architecture

Most quantitative traits are determined by a web of many interacting loci and by an array of environmental factors (Falconer and Mackay 1996). The traditional *polygenic* theory of quantitative traits (Mather 1943) envisaged a fairly large number of loci, each with relatively small and equal effects, acting in a largely additive way. Over the years it has indeed been observed that a quantitative trait may display complicated genetic architecture (Mackay 1996, 2001), expressed as

(1) It may be controlled by a fairly large number of loci; for example, of the order of 50, according to the work of Shrimpton and Robertson (1988a,b);
(2) Genes act in ways which may be additive, dominance, epistatic with other genes, and interactive with environmental factors;
(3) The magnitude of the effect produced by each locus can vary considerably;
(4) The same genes may affect different phenotypic traits through pleiotropic effects;

(5) The genes affecting the trait may be distributed over the genome at random or in a certain pattern.

With a deep use of genetic mapping to analyze quantitative traits, increasing evidence has been observed for the third point, which suggests that typically a small number of loci account for a very large fraction of the variation in the trait. For this reason, the traditional polygenic model may be replaced by a new *oligogenic* model in which a small number of major genes each with a large effect, combined with many minor genes each with a small effect, determine the genetic variation of a quantitative trait (see Mackay 1996 for an excellent review).

1.7.7 The Estimation of Gene Number

The actual number of genes that control a quantitative trait is one of the most important elements for the genetic architecture of the trait. Gene number can be estimated by a biometrical approach, although it depends on some critical assumptions (Lande 1981; Lynch and Walsh 1998). The number of genes estimated by this approach basically reflects the effective number of genes that contribute a major part of genetic variation of a trait. The most widely used approach for estimating gene number is based on the phenotypic means and variances of two parental lines and their hybrids, i.e., F_1, F_2 and backcrosses. The biometrical approach for the enumeration of effective genes was first proposed by Castle (1921).

Suppose there are two contrasting parental lines, one (P_1) being homozygous for all increasing alleles and the second (P_2) being homozygous for all decreasing alleles. These two lines are crossed to generate the F_1 and F_2. There are a total of unlinked m_e effective genes each with the same effect (a) that is purely additive. The mean phenotype of the P_1 and P_2 line can be written, respectively, as

$$\mu_{P_1} = \mu + \sum_{i=1}^{m_e} a = \mu + m_e a,$$
$$\mu_{P_1} = \mu - \sum_{i=1}^{m_e} a = \mu - m_e a,$$

whose difference is

$$\Delta = \mu_{P_1} - \mu_{P_2} = 2m_e a, \tag{1.14}$$

with the overall mean μ being canceled. Based on equation (1.7), the genetic variance of the F_2 is rewritten as

$$V_G = \frac{1}{2}\sum_{i=1}^{m_e} a^2 = \frac{1}{2} m_e a^2, \tag{1.15}$$

under the assumptions as mentioned above. Combining equations (1.14) and (1.15), we obtain the Castle-Wright estimator of gene number as

$$\hat{m}_e = \frac{\Delta^2}{8\mathrm{V_G}}, \tag{1.16}$$

where $\mathrm{V_G}$ is estimated by equation (1.12) or (1.13). The sampling variance of $\hat{n}_e$ can be approximated by

$$\mathrm{Var}(\hat{m}_e) = \hat{m}_e^2 \left[\frac{4(\mathrm{V_{P_1}} + \mathrm{V_{P_2}})}{\Delta^2} + \frac{\mathrm{Var}(\mathrm{V_G})}{\mathrm{V_G^2}} \right], \tag{1.17}$$

where

$$\mathrm{Var}(\mathrm{V_G}) = \frac{2\mathrm{V_{F_2}^2}}{n_{\mathrm{F_2}} + 2} + \frac{2\mathrm{V_{F_1}^2}}{n_{\mathrm{F_1}} + 2}$$

with $n_{\mathrm{F_2}}$ and $n_{\mathrm{F_1}}$ being the sample sizes, if equation (1.12) is used.

After the Castle-Wright estimator, several studies were pursued to improve the estimation of gene number. Lande (1981) generalized the Castle-Wright estimator for use with outcrossing populations. Zeng et al. (1990) and Zeng (1992) relaxed some of the critical assumptions, including unlinkage and equal additive effect, used for the Castle-Wright estimator. Epistatic effects between different genes were considered in Wu (1996) who extended gene enumeration to estimate a more complete picture of genetic architecture. In particular, Wu's model allows for the estimation of more genetic parameters by including multiple generations, $\mathrm{P_1}$, $\mathrm{P_2}$, $\mathrm{F_1}$, $\mathrm{F_2}$ and backcrosses, in the same experiment. Generally speaking, use of biometrical approaches for the estimation of gene number has been limited in practice, despite their significance in helping to understand general quantitative genetic theory. A more precise approach for gene enumeration is based on genetic mapping with molecular markers in which the association between markers and phenotypic variation is analyzed and tested by statistical models (Lander and Botstein 1989).

1.8 Molecular Genetics

Molecular genetics applied to linkage analysis is concerned with genetic marker technologies. Molecular genetic markers are readily assayed phenotypes that have a direct 1:1 correspondence with DNA sequence variation at a specific location in the genome. In principle, the assay for a genetic marker is not affected by environmental factors. Genetic markers are DNA sequence polymorphisms that show Mendelian inheritance. For genome mapping, the ideal genetic marker is codominant, multiallelic, and hypervariable (i.e., segregates in almost every family). However, some dominant markers are also very useful and powerful in particular situations.

Molecular markers have many different types. Restriction fragment length polymorphisms (RFLPs) were the first genetic markers that were widely used for genomic mapping and population studies. RFLP markers are obtained by using restriction endonucleases to precisely cleave a genomic DNA fragment containing a particular gene sequence. If two organisms differ in the distance between sites of cleavage of a particular restriction endonuclease, they will produce different lengths of the fragments when the DNA is digested with a restriction enzyme. The fragments can then