1 Introduction

Around 1900, scientists tried to understand the relationship between the inheritance of simple traits and observations of meiotic cell division under a microscope. It was the time that Mendel's laws on the inheritance of traits (Mendel, 1865) were rediscovered. Mendel did not have any idea of the biological mechanisms underlying his laws. However, some 35 years later, after studying Boveri's 1902 paper, Sutton (1902) realized that chromosomes and their behaviour in meiotic cell division could very well explain Mendel's results. However, in the many experimental crosses carried out to study the simultaneous inheritance of two traits, occasionally large deviations from Mendel's Law of Independent Assortment were observed. Bateson *et al.* (1905) described these deviations in terms of coupling of the heritable factors determining the traits. In subsequent work by Morgan and others, it became clear that heritable factors could be grouped with respect to the law of independent assortment: if two factors belonged to the same group, then their inheritance showed some interdependence; otherwise, the law would hold. In other words, they started realizing that these groups corresponded with chromosomes. By 1911, Morgan showed the possibility of recombination between factors lying on the same chromosome (Morgan, 1911a). Morgan assumed that this was due to an interchange (as he called the crossing over) between homologous chromosomes during meiosis. This corresponded very well with the detailed cytological observations of Janssens (1909). Later, Morgan (1911b) suggested that the heritable factors should be located in a linear fashion on the chromosomes and that the degree of coupling between traits would depend on the distance between the factors on the chromosomes. Sturtevant (1913), a student of Morgan, tested the hypothesis that the proportion of crossovers (as he called the recombinants) could be used as an index of the distance between two factors. He argued that, after determining distances from A to B and from B to C, it should be possible to predict the distance from A to C and to determine the order of the factors on the chromosome. Sturtevant successfully analysed six factors located on the sex chromosome of *Drosophila* and thus became the first ever person producing a genetic map. With his work, the chromosome theory of inheritance became really established.

What exactly is a genetic map? It is a representation of the relative positions of genes and genetic markers on the chromosomes of a biological species. Genes are functional stretches of DNA, whilst genetic markers represent anonymous, non-functional stretches of DNA. Both have fixed positions on the chromosomes, which is why together they are called loci throughout this book. The word 'loci' is the plural of the Latin word 'locus', meaning position. The genetic map of a chromosome (or part of it) is one-dimensional, reflecting the chromosome's linear structure. Distances on a genetic map are measured in units called Morgans, named in honour of the geneticist T. H. Morgan (1866–1945); in practice, centiMorgans are usually used, abbreviated to cM.

Genetic mapping is the procedure for the construction of a genetic map. Genetic linkage analysis is a more general term for any study on the co-segregation of two or more loci. Because genetic maps are always determined with linkage analysis, they are also called linkage maps and even genetic linkage maps. Since Sturtevant's first map, countless genetic maps of many biological species have been produced. In the early years, the goal was to study the chromosomal theory of inheritance, whilst later, relationships in the inheritance of traits became the prime objective. For a long time, mapped traits were simple, morphological characters, whilst later, physiological traits and isozymes were added. One of the major limitations was the limited level of polymorphism encountered with this kind of traits. Without polymorphism, segregation cannot be observed and observations on segregation are a prerequisite for mapping a trait. The development of molecular biology in the 1970s and more specifically the advent of molecular genetic markers started an enormous expansion of genetic linkage mapping. With molecular techniques, we now have an abundance of polymorphic genetic markers. Presently, genetic linkage maps play a prominent role in various fields of fundamental and applied genetic research, for instance, map-based cloning of genes, whole-genome sequencing, quantitative trait loci (QTL) analysis and marker-assisted breeding.

The subject of this book is limited to linkage analysis of experimental populations of species with a regular disomic inheritance. In diploid species, each chromosome has two copies, one inherited from each of the two parents. Diploid species have a disomic inheritance, which is described in Chapter 2 on meiosis. In diploid species, meiosis is the cell division process that produces haploid gametes, which carry a single copy of each chromosome. As mentioned above, segregation is a prerequisite of linkage analysis. If an individual carries different alleles, i.e. variants, at a locus on the two corresponding, homologous chromosomes, this locus will segregate in the offspring of this individual. The essence of genetic linkage analysis is observing how often alleles of loci are inherited together or become exchanged due to crossovers in meiosis. Because the exchange due to a crossover results in new combinations of alleles, this phenomenon is called genetic

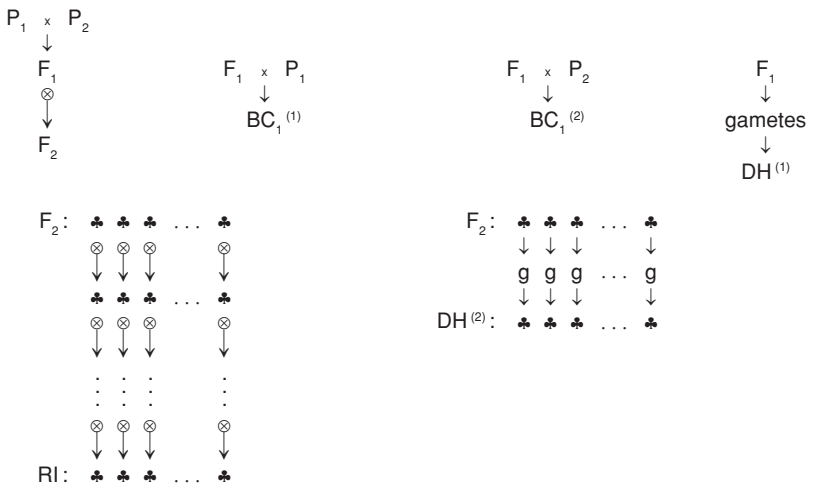

Fig. 1.1

Common experimental population types based on fully homozygous parents, P_1 and P_2. Crossing the parents results in the heterozygous F_1. The F_1 can be selfed (⊗) to obtain an F_2 population. The F_1 can be crossed back to each of the parents, resulting in a first-generation backcross population: $BC_1{}^{(1)}$ or $BC_1{}^{(2)}$. A special treatment of the haploid gametes (g) of the F_1 or of each F_2 individual results in a doubled haploid population: DH $^{(1)}$ or DH $^{(2)}$. Repeated selfing of each individual by single-seed descent starting from the F_2 results in a recombinant inbred line (RI) population. The symbol ♣ represents an individual; the ellipsis (. . .) indicates repetition.

recombination. Two loci are said to be closely linked when there is a strong tendency that the original combinations of alleles are transmitted together to the next generation, i.e. very little genetic recombination occurs. The strength of the linkage is inversely related to the distance between the loci on the chromosomes.

In order to study how often alleles of loci are inherited together or become exchanged by genetic recombination, we have to create segregating progenies (hereafter often referred to as populations). For obvious reasons, this is not an option in human linkage studies; here, segregation must be observed in (large) sets of existing pedigrees. The same applies to linkage studies in certain (livestock) animals. This book is limited to linkage analysis in experimental populations.

An individual carrying different alleles at a certain part of its loci, i.e. it is heterozygous at these loci, will produce different gametes. Therefore, its progeny will consist of different genotypes, in other words, it will reveal segregation for these loci. Figure 1.1 shows a scheme with the most common types of experimental population derived from fully homozygous parent lines, also known as inbred or pure lines. The alleles of a homozygous locus are identical at the homologous chromosomes. A pure line is homozygous over its entire genome. Crossing two distinct pure lines will create an F_1 that is heterozygous at all loci at which the

two parent lines differ in alleles. The heterozygous F_1 is the starting point for the common segregating population types, such as the first-generation backcross (BC_1) and the F_2. For certain species, it is possible to produce so-called doubled haploids (DH). With a special treatment of the male or female gametes, the haploid set of chromosomes of a gamete is duplicated and subsequently the gamete develops into a diploid, fully homozygous individual. Another common population type is the family of recombinant inbred lines (RI or RIL), which is obtained by repeated inbreeding starting from the F_2 using a process called single-seed descent. Each type of population has its advantages and disadvantages. For instance, in a RIL family and in a DH population, all individuals are homozygous at practically all loci. Because of this, they can be reproduced indefinitely by seeds without changing their genotypes. Such populations are called immortalized populations.

After a segregating population has been created, the next step is to observe the genotypes of the markers, genes and traits of each individual. As already mentioned, genetic linkage maps were originally constructed using mainly morphological markers. Such markers are limited in number; moreover, combinations of alleles at multiple loci may have severe side effects on the resulting phenotypes. Currently, several types of molecular genetic markers are employed. The unique characteristics of the DNA molecule lie at the basis of all molecular marker techniques. Enzymatic cutting of DNA into fragments (also called restriction digestion), electrophoretic separation of fragments, hybridization, ligation and DNA synthesis are the biochemical techniques involved. Molecular markers visualize variation in coding and non-coding parts of the DNA. Molecular variants can be followed genetically, just like the alleles of segregating genes. In molecular size, a molecular genetic marker may represent just a single DNA nucleotide up to a stretch of thousands of nucleotides. As a chromosome consists of billions of nucleotides, molecular markers may be considered as points on chromosomes. Currently, the number of markers employed in a mapping experiment has grown into the thousands for a whole genome. Recent developments in single-nucleotide polymorphism (SNP) technologies are increasing this number dramatically, into the tens of thousands.

With respect to linkage analysis per se, the only difference between the various molecular marker techniques is whether it is possible to identify all possible genotypes of a segregating population. In the case of dominance, the heterozygote cannot be distinguished from one of the homozygotes. Usually, this is caused by the presence of so-called null alleles, which are alleles representing the absence of any (observable) molecular variant. In this book, we do not pay attention to the various molecular marker techniques. However, we want to emphasize that it is important to have at least a basic knowledge of the techniques employed in order to be able to cope with possible problems encountered in linkage analysis.

When all observations on the markers, genes and traits are recorded, the actual linkage analysis begins. As mentioned above, the essence of genetic linkage analysis is observing how often the alleles of loci are inherited together or are exchanged due to genetic recombination. The rate of recombination between two loci is a measure of the distance apart of these loci. Measuring these rates for all pairs of loci is the starting point for constructing the linkage map. In Chapter 3, we deal with the estimation of the recombination frequencies in various common situations. For this purpose, we introduce the method of maximum likelihood, which is a general statistical method for the estimation of parameters.

A genetic linkage map describes the relative positions of genes and genetic markers on the chromosomes. Loci on the same chromosome have a linear arrangement, whilst loci on different (non-homologous) chromosomes are inherited independently. Consequently, no linear relationships exist between the different chromosomes. This is why genetic linkage maps are estimated separately for each chromosome. Determining which loci belong to the same chromosome is therefore a necessary step in the construction of linkage maps. Loci on the same chromosome are physically linked, whereas those on different chromosomes are physically unlinked. Genetic linkage is the phenomenon where traits have a tendency to be inherited together. Due to the independent assortment of the chromosomes in the meiosis, genetic linkage is the result of physical linkage. Determining whether two loci are genetically linked is the starting point of clustering loci into groups. Sets of genetically linked loci are called linkage groups. Chapter 4 tackles the methods for obtaining the linkage groups and describes how to deal with possible complications.

A linkage map is given in linear distance units, whilst the linkage analysis begins with determining recombination frequencies. Therefore, Chapter 5 starts by describing the relationship between recombination frequencies and map distances. Distances can be obtained from recombination frequencies by applying a so-called mapping function. Chapter 5 continues with the description of two methods that can be used to calculate a map for a given order: linear regression and maximum likelihood. The former method is applied to map distances as obtained from the recombination frequencies with a mapping function, whilst the latter is applied to the original genotype observations. The maximum likelihood method estimates recombination frequencies. Because the data of all loci are taken into account, these estimates are often called multipoint estimates of recombination frequencies, in contrast to two-point or pairwise estimates, which are based on pairs of loci only. The multipoint recombination frequencies are subsequently translated into map distances using a mapping function.

Although calculating a genetic linkage map for a given order may be straightforward, the correct order is usually unknown. The order cannot be observed directly

from the segregation data, but must be inferred. To do this, we calculate a map for a certain given order and next determine some criterion that expresses the quality of the result. We may repeat this for many possible orders, and identify the best order according to the criterion. In Chapter 6, we present several criteria for the evaluation of maps. Applying these to a small example dataset illustrates that they generally all point to the same order as the best.

As the number of markers increases, the number of possible map orders increases exponentially, so that quite soon it becomes infeasible to evaluate all possible orders. Efficient optimization techniques are required to obtain the best map for a given set of data, thereby circumventing the need to evaluate all possible orders. The techniques must not only be efficient in the sense of computational speed. Fast but poor methods may result in orders that are optimal for parts of the linkage group only. Therefore, the techniques must also be successful in finding orders that are close to the optimal order for the entire linkage group. Chapter 7 is dedicated to optimization algorithms that have been applied to the locus-ordering problem.

The genetic circumstances found in outbreeding species are more complicated than in inbreeding species: (i) more than just two alleles per locus may be present, and (ii) the linkage phases may vary across the loci and between the parents of an experimental cross. Consequently, linkage analysis of outbreeding species actually concerns the analysis of the segregation in two distinct meioses. The two meioses can be analysed separately, which is called the two-way pseudo-testcross strategy. However, it can also be done simultaneously. Both approaches will result in two separate maps, which may be combined into one as an average for both meioses. Chapter 8 explains in detail the genetic situations encountered in outbreeding species and describes the linkage analysis of a full-sib family of an outbreeding species.

We dedicate the final chapter to practical aspects of genetic linkage mapping. The analysis methods presented in this book are based on current knowledge of meiosis. This knowledge is converted into a mathematical model. Experimental data are analysed according to this model resulting in a map. Essentially, a map is used for prediction purposes: in map-based cloning, for indirect selection in breeding or in QTL analysis. Current experience shows that linkage mapping is a very reliable tool if the data to be analysed are of high quality. In practice, however, experimental data may suffer from missing observations and errors. Segregation may occur in a non-Mendelian way, i.e. not according to our model of meiosis. Consequently, the data may not fit the model (or, vice versa, the model is not in accordance with the data). Therefore, prior to the linkage mapping, the data should be inspected carefully in order to assess the quality. Errors may be random or systematic; if they are systematic, any poor-quality marker or individual should be identified and removed. If the poor-quality observations concern the gene of

Table 1.1. Genetic key figures (roughly rounded) for some selected species. *N*, haploid number of chromosomes; TML, total map length (cM); CML, average chromosome map length (cM); bp, number of DNA base pairs ($\times 10^9$); bp/TML, average number of base pairs per cM ($\times 10^6$).

Species	*N*	TML	CML	bp	bp/TML
Arabidopsis	5	600	120	0.15	0.25
Barley	7	1000	140	5	5
Caenorhabditis elegans	6[a]	500	85	0.1	0.2
Lily	12	1000	85	40	40
Maize	10	1700	170	3	1.8
Man	23[a]	3500	150	3	0.9
Onion	8	700	90	10	14
Table mushroom	13	1200	90	0.03	0.03
Tomato	12	1400	120	0.9	0.65

[a] Including the sex chromosome.

interest, removal is not an option, so the quality of the observations will need to be improved. Sometimes, systematic errors may not surface until after a linkage map has been calculated. If so, poor-quality markers or individuals should be identified and removed and the map recalculated. Thus, the mapping procedure can be characterized as an iterative process of pre-mapping diagnostics, actual mapping and post-mapping diagnostics.

What is the relationship between a genetic linkage map and a physical map? A genetic linkage map concerns genetic recombination probabilities, whilst a physical map is about base pairs (or nucleotides) and sequences. Table 1.1 lists some genetic key figures for a few species. You can see that there is quite some variation in the haploid number of chromosomes, but that the average length of the genetic linkage map of a chromosome has a limited range, with an overall average of roughly 100 cM. It usually varies between 50 and 200 cM among and within species. At the same time, a considerable amount of variation is present in the number of base pairs across species. Moreover, a more than 1000-fold range in the number of base pairs is present in a centiMorgan map length. In other words, across species there is no constant one-to-one relationship between the linkage and the physical map. Furthermore, no such constant relationship exists within species: because recombination itself is a biological process, you may expect to

find variation among individuals within the same species. Finally, there are so-called hot and cold spots of recombination: on certain physical locations on the chromosomes, relatively many or only a few recombination events take place. This means that, even within an individual, a large amount of variation exists in the numbers of base pairs for each centiMorgan map length.

Linkage mapping requires a basic understanding of genetics, molecular biology, probability, statistics and optimization techniques. The goal of this textbook is to provide the reader with an adequate level of knowledge on these subjects without going into too much mathematical detail. This should enable him or her to produce linkage maps of experimental populations (with the help of computer software). In addition, he or she should be able to recognize and subsequently deal with various problematic situations that may be encountered in practice.

References

Bateson, W., Saunders, E. E. & Punnett, E. C. (1905). Experimental studies in the physiology of heredity. *Reports to the Evolution Committee of the Royal Society. Report II*. http://www.archive.org.

Boveri, T. (1902). Über mehrpolige Mitosen als Mittel zur Analyse des Zellkerns. *Verhandlungen der physicalisch-medizinischen Gesselschaft zu Würzburg. Neue Folge*, 35, 67–90.

Janssens, F. A. (1909). La théorie de la chiasmatypie. Nouvelle interprétation des cinèses de maturation. *La Cellule*, 25, 389–411. http://archive.org/stream/lacellule25lier. Translated to English: *Genetics*, 191 (2012), 319–346.

Mendel, G. J. (1865). *Versuche über Pflanzenhybriden*. Verh. des Naturf. Vereins, Brünn. IV. Band. Abhandlungen 1865, Brünn, 1866. S. 3–47. http://www.biologie.uni-hamburg.de/b-online/e08/08a.htm and http://www.mendelweb.org.

Morgan, T. H. (1911a). The application of the conception of pure lines to sex-limited inheritance and to sexual dimorphism. *American Naturalist*, 45, 65–78. http://www.jstor.org/stable/2455465.

Morgan, T. H. (1911b). Random segregation versus coupling in Mendelian inheritance. *Science*, 35, 384. http://www.jstor.org/stable/1638198.

Sturtevant, A. H. (1913). The linear arrangement of six sex-linked factors in *Drosophila*, as shown by their mode of association. *Journal of Experimental Zoology*, 14, 43–59. http://www.esp.org/timeline.

Sutton, W. S. (1902). On the morphology of the chromosome group in *Brachystola magna*. *Biological Bulletin*, 4, 24–39. http://www.esp.org/timeline.