

Chapter 2

A Primer in Genetics

2.1 Basic Biology

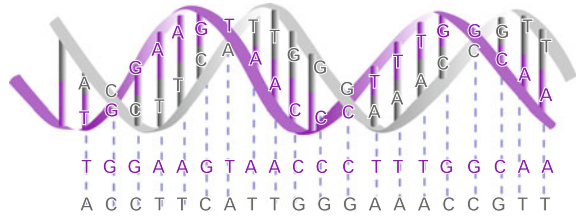
2.1.1 Phenotypes and Genotypes

A **phenotype** is any observable characteristic of an organism. Phenotypes of interest could be, for example, height, weight, blood pressure, blood type, eye color, disease status, the size of a plant's fruits, or the amount of milk given by a cow. Typically, one observes quite a large amount of variety in phenotypes between individuals of the same species. Phenotypes are influenced by both genetic and environmental factors. A great proportion of current biological research consists of trying to get a better understanding of the genetic factors involved.

In eukaryotes (organisms composed of cells with a nucleus and organelles), including plants, animals, or fungi, most of the genetic material is contained in the cell nucleus. This material is organized in **deoxyribonucleic acid (DNA)** structures called **chromosomes**. DNA consists of two long polymers of simple units called nucleotides. One element of a nucleotide is the so-called nucleobase (nitrogenous base). There are four primary DNA-bases: cytosine, guanine, adenine, and thymine, abbreviated as C, G, A, and T, respectively. Pairs of DNA strands are joined together by hydrogen bonds between complementary bases: A with T, and C with G. Therefore, the sequence of nucleotides in one strand can be determined by the sequence of nucleotides in the other (complementary) strand. The backbone of a DNA strand is made from phosphates and sugars joined by ester bonds between the third and fifth carbon atoms of adjacent sugar rings. The corresponding ends of DNA strands are called the 5' (five prime) and 3' (three prime) ends. Such a pair of DNA strands are orientated in opposite directions, 3'–5' and 5'–3'. Therefore, they are called **antiparallel**.

A pair of DNA strands form a structure known as the **double helix**, illustrated in Fig. 2.1. However, for the purpose of many statistical and bioinformatical analyses, chromosomes are simply represented as sequences, where each element is the letter

Fig. 2.1 An illustration of the double helix structure and two antiparallel sequences of nucleobases



corresponding to the nuclear base (C, G, A, or T) at the corresponding position in one of the strands.

In the process of **transcription**, some sections of DNA, called **genes**, are transcribed into complementary copies of ribonucleic acid (RNA). Since RNA is single stranded, only one strand of DNA is used in the transcription process. The resulting RNA strand is complementary and antiparallel to the “parental” DNA strand, with thymine (T) being replaced by uracil (U). As a result, the RNA sequence is identical (except for T being replaced by U) to the complementary sequence of the parental DNA strand.

If a gene encodes a protein, then the resulting messenger RNA (mRNA) is used to create that protein through the process of **translation**. Proteins can be viewed as chains of amino acids, where certain triplets of mRNA are translated into specific amino acids. In eukaryotes there is a further modification of RNA between the processes of transcription and translation, which is called **splicing**. Here, parts of the RNA, so-called **introns**, are removed and the remaining parts, called **exons**, become attached to each other. After splicing, the mRNA consists of a sequence of triplets, which directly translate into the amino acids forming the protein expressed by the gene in question.

The DNA sequence corresponding to an mRNA sequence is called a **sense** strand. Thus, as explained above, a sense DNA sequence is complementary to the corresponding parental (**antisense**) DNA sequence. Both strands of DNA can contain sense and antisense sequences. Antisense RNA sequences are also produced, but their function is not yet well known. Proteins, as well as functional RNA chains, created via transcription and translation play an important role in biological systems and influence many phenotypes.

The process of gene expression depends not only on the coding region, but also on the regulatory sequences that direct and regulate the synthesis of gene products. **Cis-regulatory** sequences are located in the close vicinity of the corresponding gene. They are typically binding sites for transcription factors (usually proteins), which regulate gene expression. **Trans-regulatory** elements are DNA sequences that encode these **transcription factors** and are not necessarily close to the gene in question. They may even be found on different chromosomes.

The DNA sequences of different individuals from a given species are almost identical. For example, in humans 99.9 % of all DNA-bases match. However, there still exist a large number of **polymorphic** loci, at which differences between individuals from a given species can be observed. The variants observed at such a locus

are called **alleles**, where the most prominent examples of such genetic variation are **single nucleotide polymorphisms** (SNPs) and **copy number variations** (CNVs). SNPs refer to specific positions in a chromosome where different nucleobases are observed, the result of a so-called point mutation. Copy number variation refers to relatively long stretches of DNA which are repeated a different number of times in various individuals. In particular, insertions, deletions, and duplications of DNA stretches are classified as CNVs. If the DNA section corresponding to a CNV includes a gene, it will result in different gene expression patterns. **Microsatellites** are also classical examples of genetic polymorphisms, where very short DNA patterns are repeated a number of times, and the number of repetitions varies between individuals.

The number of **homologous** chromosomes, which at a given locus contain genes corresponding to the same characteristic, varies between different species. **Haploid** organisms, such as male bees, wasps, and ants, have just one set of chromosomes (i.e., just one copy of each gene). The majority of all animals, including humans, are **diploid**, i.e., they have two sets of chromosomes, one set inherited from each parent. In diploid organisms an individual's **genotype** at a given locus is defined by the pair of alleles residing at this locus on the two homologous chromosomes. For example, consider a biallelic locus with alleles A and a. Then there exist three possible genotypes: AA, Aa, and aa. An individual carrying two identical alleles at a given locus is called **homozygous** at this locus, whereas an individual with two different alleles is **heterozygous**. There also exist many organisms which are **polyploid**, meaning that they have more than two homologous chromosomes. Polyploid organisms are common among plants, e.g., the potato, cabbage, strawberry, and apple. In this book, we will mainly focus on methods for localizing genes in diploid organisms.

A **haplotype** is an ordered sequence of nucleobases appearing on the same chromosome. For example, a haploid organism inherits a maternal haplotype and a paternal haplotype, which together define the genotypes at the corresponding loci. When an individual is genotyped, generally we do not know which parent each allele came from. In this case, we say that the genotypes are unphased. Hence, it might be necessary to infer the haplotypes from the genotype data (in other words, determine the phase). One of the most popular algorithms for phasing is FASTPHASE [113], which applies maximum likelihood methods to predict haplotypes. In this book, we will mainly focus on statistical methods which make use of genotype data, although many of the statistical methods described in Chap. 5 can be extended to phased haplotype data. For illustrative purposes, Table 2.1 gives a simple example of unphased genotypes at 10 markers, and two phased haplotypes corresponding to these genotypes.

Table 2.1 Unphased genotypes and phased haplotypes for 10 markers

Unphased	aA	BB	cC	dD	ee	ff	gG	hH	iI	JJ
From father	A	B	c	d	e	f	G	H	i	J
From mother	a	B	C	D	e	f	g	H	I	J

The genetic information defining gender is typically contained in **sex chromosomes**. Among diploid organisms, the XX/XY sex-determination system is the most common. In this system, females have two sex chromosomes of the same kind (XX), while males have two distinct sex chromosomes (XY). The X and Y sex chromosomes are different in size and shape from each other. The Y chromosome contains a gene called SRY (the sex determining region of Y) that determines maleness and can only be inherited from the father. In humans the X chromosome spans more than 153 million base pairs coding for approximately 2000 genes, while the Y chromosome spans about 58 million base pairs and contains 86 genes, which code for only 23 distinct proteins. Traits that are inherited via the Y chromosome are called holandric traits. The remaining chromosomes, i.e., those not related to gender, are called **autosomes**.

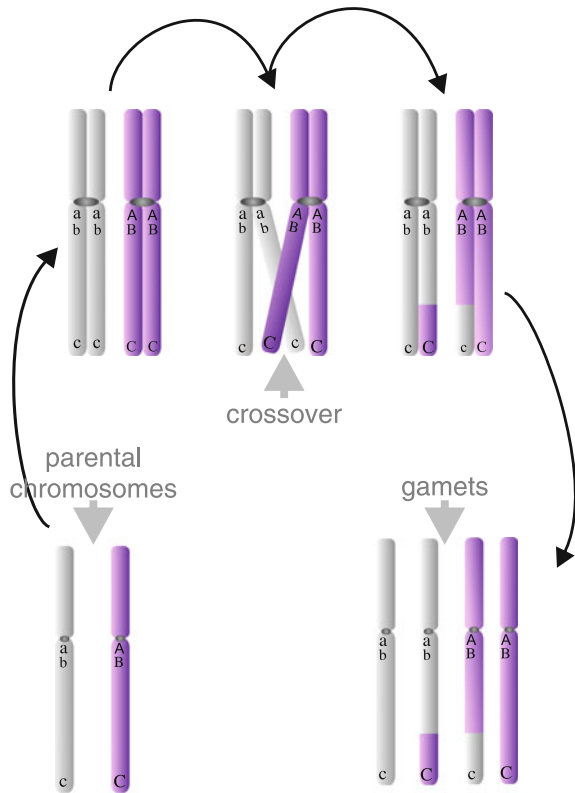
2.1.2 Meiosis and Crossover

In many organisms genetic material is passed on from parents to offspring through the process of **sexual** reproduction. **Meiosis** is the biological process via which gametes or spores are produced. In diploid organisms meiosis transforms one diploid cell into four haploid gametes. Before meiosis, the paternal and maternal copies of each chromosome duplicate themselves and form two pairs of identical **sister chromatids**, where each pair are joined together at the centromere. Then the maternal and paternal homologues pair with each other. Occasionally, genetic material is exchanged between a paternal and maternal chromosome in events called **crossover**, where matching regions of both chromosomes break and then reconnect with the other chromosome. In the final stage of division, the cell is divided into four gametes, each containing one homologous chromatid. After crossover, such a gamete may contain genetic material from both maternal and paternal homologues. If the genetic material at two loci comes from different parents, then we say that **recombination** has occurred between these two loci. If two loci reside on different chromosomes, then the probability of recombination between them is equal to $1/2$. If two loci reside on the same chromosome, then recombination results from an odd number of crossovers. Recombination increases the genetic diversity of a population. An illustration of meiosis is given in Fig. 2.2.

2.1.3 Genetic Distance

The physical distance between two loci is often expressed in terms of the number of nucleotide bases lying between them. However, for the purpose of gene mapping, scientists normally use genetic maps, which express distance in terms of probabilistic units. Genetic distance is often measured in **Morgans**. A distance of 1 Morgan means that the expected number of crossovers between two loci in a single meiosis

Fig. 2.2 A graphical illustration of meiosis



is equal to 1. There exist several **mapping functions**, which relate the probability of recombination to the genetic distance.

2.1.4 The Haldane Mapping Function

In practice, the **Haldane function** is the most frequently used mapping function. It assumes a **lack of interference**, which means that crossover events occur independently of each other. Under this assumption, the number of crossovers on a given piece of a chromosome can be modeled by the Poisson distribution (see Sects. 6.3.1 and 6.5.2).

Let the distance between two loci be equal to d Morgans. Then the number X of crossovers between these loci has a Poisson distribution with mean d ,

$$P(X = k) = \frac{d^k}{k!} \exp(-d), \quad \text{for } k \in \{0, 1, 2, \dots\}.$$

Thus the probability of recombination between these two loci, r , is given by the formula

$$\begin{aligned} r &= \sum_{k=0}^{\infty} P(X = 2k + 1) = \sum_{k=0}^{\infty} \frac{d^{2k+1} \exp(-d)}{(2k + 1)!} \\ &= \sinh(d) \exp(-d) = \frac{e^d - e^{-d}}{2} e^{-d} = \frac{1}{2}(1 - e^{-2d}). \end{aligned} \quad (2.1)$$

Note that according to this formula $0 \leq r \leq 1/2$, where $r = 0$ corresponds to a genetic distance of 0, and $r = 1/2$ to an infinite genetic distance. As the genetic distance increases, the recombination rate converges rapidly to $1/2$, i.e., for large d the recombination rate is very close to $1/2$. For loci on different chromosomes, one usually defines $r = 1/2$. When $r < 1/2$, then we say that two loci are linked (i.e., lie on the same chromosome).

Solving Eq. 2.1 for d yields the

Haldane mapping function:

Let r be the probability of recombination between two loci and d be the genetic distance between these loci (in M). Under the assumption of no interference, it follows that $d = H(r)$, where

$$H(r) := -(1/2) \ln(1 - 2r) \quad (2.2)$$

is the Haldane mapping function.

2.1.5 Interference and Other Mapping Functions

Consider three genetic loci at positions $L_1 < L_2 < L_3$ on the same chromosome, and denote by R_{ij} the event that recombination occurs between loci i and j . Then it is immediately clear that

$$R_{13} = (R_{12} \cup R_{23}) \setminus (R_{12} \cap R_{23}).$$

If the events R_{12} and R_{23} are independent, then the probabilities r_{ij} of recombination between loci i and j satisfy

$$r_{13} = r_{12} + r_{23} - 2r_{12}r_{23}.$$

This equality is fundamental to defining the Poisson process underlying the derivation of the Haldane mapping function. However, in practice, one often observes that

the occurrence of crossover at one locus reduces the chance of crossover at neighboring loci. Due to this kind of **interference**, the probability of a double crossover in a given interval is smaller than that estimated by applying the Poisson process and, consequently, the probability of recombination between neighboring loci is larger than the estimate provided by the Haldane function. In general, the recombination fractions will satisfy equation

$$r_{13} = r_{12} + r_{23} - 2Cr_{12}r_{23},$$

where $C \in [0, 1]$ is the coefficient of **coincidence** and $I = 1 - C$ is the coefficient of interference. In the case $I = 0$, there is no interference, resulting in the genetic distance being described by the Haldane function (2.2). The other extreme situation, $I = 1$, corresponds to complete interference, which eliminates the possibility of more than one crossover on any chromosome. In this case, the genetic distance d and the probability of recombination r are connected via the simple

Morgan mapping function:

In the case of complete interference, one has

$$d = M(r) := r.$$

Note that according to the assumption of complete interference, the expected number of crossovers on each chromosome cannot exceed 1, which limits the maximal length of a chromosome to 1 Morgan.

Another popular mapping function assumes $I = 0.5$. This intermediate choice of the interference parameter gives rise to the

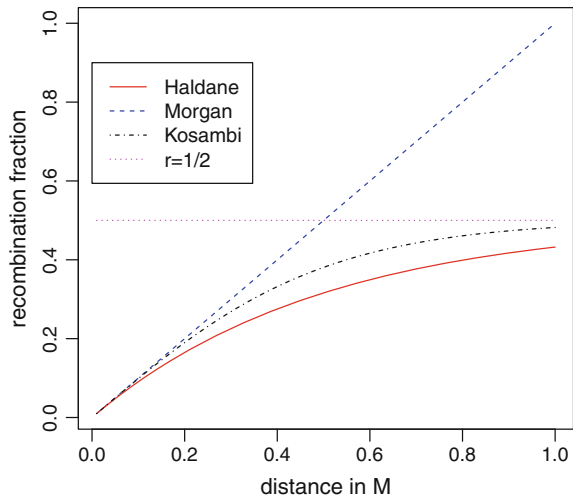
Kosambi mapping function:

$$d = K(r) = \frac{1}{4} \ln \left[\frac{1 + 2r}{1 - 2r} \right].$$

As in the case of the Haldane mapping function, the recombination fraction $1/2$ corresponds to an infinitely long chromosome. However, for any given genetic distance, the recombination fraction according to the Kosambi function exceeds the recombination fraction given by the Haldane function.

The relationship between the recombination fraction and the genetic distance according to the three mapping functions discussed above is presented in Fig. 2.3.

Fig. 2.3 The relationship between the recombination fraction and the genetic distance according to different mapping functions



2.1.6 Markers and Genetic Maps

In Sect. 2.1.1 we briefly mentioned several possible types of genetic variation between individuals of the same species, such as SNPs or CNVs. Such polymorphisms are of primary importance in a large number of genetic studies. A polymorphic piece of a DNA sequence with known location on a chromosome serves as a **genetic marker**. Historically, the genotype at the earliest known genetic markers was determined by observing the corresponding phenotypic trait, e.g., blood group or the color of flowers. Today, a vast number of genetic markers are available, due to modern sequencing techniques. Markers are used to construct genetic maps and as reference points, to decide which parts of the genome have some influence on a given phenotypic trait.

There exist many experimental techniques for genotyping markers. Among the most popular classical systems of markers, one should mention the following: restriction fragment length polymorphisms (RFLPs), randomly amplified polymorphic DNA (RAPDs), and short tandem repeats (STRs), which are also known as microsatellites. The alleles at microsatellite markers differ in the number of replications of short (1–6 base pairs) sequences of DNA. In comparison to RFLPs and RAPDs, microsatellites have a substantially greater number of alleles and are more useful for locating genes in natural **outbred** populations. Due to the development of new DNA sequencing techniques, the determination of Copy Number Variations (CNVs) has become possible, an approach which has gained large popularity in recent years, especially in the context of human genetics.

The most popular system of markers in Genome Wide Association Studies (GWAS) is based on single nucleotide polymorphisms (SNPs). A single nucleotide polymorphism occurs when different nucleotides appear at a given single base pair within a population. As an example, consider the following two sequenced DNA

fragments from different individuals, AGCCT and AGCTT. There is a SNP at the fourth position with alleles C and T. At the vast majority of SNPs, only two alleles are possible, therefore this system of markers is essentially considered to be biallelic. The allele which is more frequent in a population is called the **reference** (or **major**) allele, while the less frequent allele is called the **variant** (or **minor**) allele. SNPs have gained large popularity in recent years, due to the development of SNP microarrays. This technology facilitates the quick and relatively cheap genotyping of several hundred thousand SNPs at the same time. Often these microarrays include a comparable number of CNVs.

2.2 Types of Study

Depending on the organism under investigation and the trait of interest, there exist a variety of experimental designs and methods to locate influential genes. When dealing with organisms which reproduce quickly and can be experimentally crossed, one can use techniques which have been specifically developed for such **experimental populations**. Based on such crosses, statistical methods for detecting a qualitative trait locus (QTL—a gene influencing a quantitative trait) are usually referred to as **QTL mapping**. It is not practical to generate experimental populations of certain species, which is particularly true for the human species, for obvious reasons. In this case, the types of studies used fall into one of the two categories: **linkage analysis**, which is based on data from families (pedigrees), and **association studies**, which are quite often performed with so-called outbred populations, which means that there are no close relatives within the study sample. The logic underlying association studies is rather similar to QTL mapping, whereas linkage analysis based on pedigree data is fundamentally different and will not be discussed in depth in this book. A good basic introduction to linkage analysis is given by [4].

2.2.1 *Crossing Experiments*

The starting point for all experimental populations are so-called **inbred lines**. These are obtained by mating only individuals from the same line over successive generations. Due to the fundamental laws of inheritance in small populations, which were described by Fisher [5], those alleles which are less frequent tend to get eliminated and after a number of generations the individuals from a given inbred line are genetically identical and homozygous at almost all loci (which means they have the same alleles at both chromosomes). These lines are typically chosen in such a way that one observes a large difference in both the phenotype of interest and the traits related to the markers between the two lines. Experimental populations are then obtained by crossing individuals from two distinct inbred lines.

Let us generically denote alleles from the first line by a , and alleles from the second line by A . The **F1 population** results from crossing individuals from both lines with each other. This generation is again genetically homogeneous, because each F1 individual has genotype Aa at each locus, which means that it has alleles from both inbred lines. In other words, each F1 individual is heterozygous at each locus. Further crosses involving F1 individuals lead to the experimental populations used in genetic studies. Depending on the exact strategy followed, one obtains different experimental designs, among which the most popular are the **backcross** design, the **intercross** design, and **recombinant inbred lines**.

2.2.1.1 Backcross Design

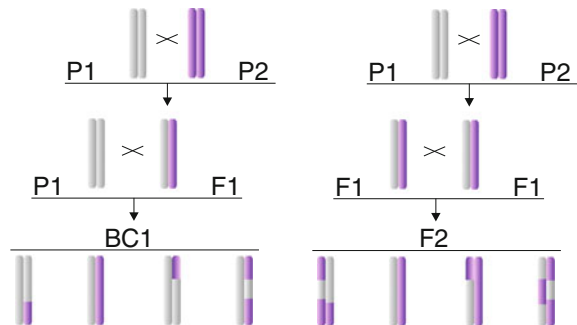
Using a backcross design, individuals from the F1 population are crossed with individuals from one of their parental inbred lines (line **P1** in Fig. 2.4). The resulting individuals form a backcross population (**BC1** in Fig. 2.4). At each genetic locus there are only two possible genotypes, either homozygous (with both alleles from the parental inbred line P1) or heterozygous. Thus, at each locus the genotype of BC1 individuals is fully determined by the allele inherited from its F1 parent.

It is often convenient to encode the genotype at a genetic locus by $X = 0$ in the case of homozygosity, and by $X = 1$ otherwise. For theoretical considerations, the state of the genotype can be thus interpreted as a random variable. Due to the crossing strategy of the backcross design, it follows that $P(X = 0) = P(X = 1) = 1/2$. Thus X is **Bernoulli distributed** with “success probability” $p = 1/2$. It immediately follows that $E(X) = 1/2$ and $\text{Var}(X) = 1/4$.

Now let us consider two genetic loci, L_1 and L_2 , residing on the same chromosome. It turns out that if L_1 and L_2 are close to each other, then it is likely that an individual from the BC1 generation will be either homozygous at both loci or heterozygous at both loci. In fact, the conditional probability that an individual is homozygous at L_1 given that it is heterozygous at L_2 is simply the probability of recombination between these two loci. In mathematical terms, this can be expressed as

$$P(X_1 = 0|X_2 = 1) = r, \quad P(X_1 = 1|X_2 = 1) = 1 - r,$$

Fig. 2.4 Descriptions of the backcross (*left*) and intercross (*right*) designs



where X_1 and X_2 denote the genotypes at L_1 and L_2 , respectively. The genetic distance between loci then directly translates into the correlation between their genotypes according to the formula

$$\text{Corr}(X_1, X_2) = 1 - 2r.$$

(2.3)

This is easily obtained by computing the covariance between the genotypes

$$\begin{aligned} \text{Cov}(X_1, X_2) &= E(X_1 X_2) - E(X_1)E(X_2) \\ &= P(X_1 = 1, X_2 = 1) - 1/4 \\ &= P(X_1 = 1|X_2 = 1)P(X_2 = 1) - 1/4 = (1 - r)/2 - 1/4 = \frac{(1 - 2r)}{4}. \end{aligned}$$

2.2.1.2 Intercross Design

Using the intercross design, often denoted as **F2**, individuals from the F1 population are crossed with each other (see Fig. 2.4). Individuals from the F2 population can have any of the three possible genotypes at each locus: AA , Aa or aa . As in the case of a backcross design, one can easily calculate the conditional probabilities of a certain genotype at locus L_2 given the genotype at locus L_1 as a function of the distance between these two loci. These probabilities are given in Table 2.2.

Using the intercross design, the classical coding of genotypes is defined by the Cockerham model, which uses two state variables X and Z for the genotype at each locus (see Table 2.3). It will be shown in Sect. 2.2.2 that the variable X is

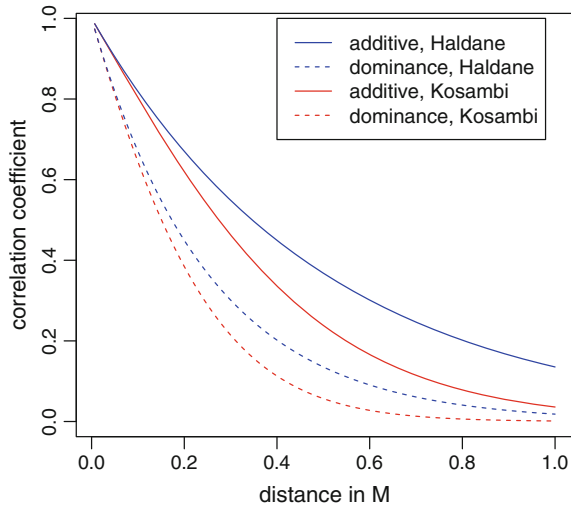
Table 2.2 Intercross (or **F2**) design: conditional probabilities of genotypes at locus L_2 given the genotype at L_1 and the probability r of recombination between these two loci

	L_2		
L_1	AA	Aa	aa
AA	$(1 - r)^2$	$2r(1 - r)$	r^2
Aa	$r(1 - r)$	$r^2 + (1 - r)^2$	$r(1 - r)$
aa	r^2	$2r(1 - r)$	$(1 - r)^2$

Table 2.3 Coding of genotypes using the Cockerham model

	Dummy	Variables
Genotype	X	Z
AA	−1	−0.5
Aa	0	0.5
aa	1	−0.5

Fig. 2.5 The correlation coefficient between variables coding for additive and dominance effects as a function of the genetic distance



associated with the additive effects of a QTL (a measure of the mean difference in the value of the trait between the two types of homozygote). Similarly, the variable Z is associated with the dominance effect (this is zero if the mean value of the trait among heterozygotes is exactly half way between the mean values in the homozygote groups, i.e., neither allele dominates the other). Based on the conditional probabilities from Table 2.2, it is a simple exercise to compute the correlation between these two dummy variables according to the genetic distance:

$$\text{Corr}(X_1, X_2) = 1 - 2r, \quad \text{Corr}(Z_1, Z_2) = 4(r - 0.5)^2, \quad (2.4)$$

$$\text{Corr}(X_1, Z_1) = \text{Corr}(X_1, Z_2) = 0,$$

where r represents the probability of recombination between the two loci.

The correlation coefficient between these state variables as a function of the genetic distance is illustrated in Fig. 2.5. It can be observed that the correlation coefficient between variables corresponding to dominance effects decays faster than the correlation coefficient between variables describing additive effects. Also, these correlation coefficients decay faster for the Kosambi mapping function than the Haldane mapping function.

2.2.1.3 Recombinant Inbred Lines

Recombinant inbred lines are obtained by the multiple crossing of close relatives or by the multiple self-fertilizing of individuals from an F₂ population. Due to the elimination of “rare” alleles, recombinant inbred lines consist of individuals who are homozygous at almost every locus, but can contain alleles from different parental

Table 2.4 Recombinant inbred lines: Probabilities R of obtaining two distinct genotypes at loci L_1 and L_2 , given the probability r of recombination between these loci

Type of inbred line	$P(X_1 \neq X_2)$
Self-fertilizing	$2r/(1 + 2r)$
X chromosome, sibling mating	$(8/3)r/(1 + 4r)$
Autosomes, sibling mating	$4r/(1 + 6r)$

inbred lines at different loci. Because only two genotypic states are possible at each locus, the methods used for the statistical analysis of recombinant inbred lines is, to a certain extent, similar to the analysis of a backcross design. However, the correlation between genetic loci is different for recombinant inbred lines. Let r be the recombination fraction between loci L_1 and L_2 . According to [7] (see also [1]), the probabilities of obtaining two distinct genotypes X_1 and X_2 at these locations are given in Table 2.4. For the backcross design, the probability of observing two different genotypes is simply given by r , which is smaller than for any recombinant inbred line. As a result, the backcross design needs less markers per chromosome to detect a QTL, but it also gives less precision with respect to the exact location.

Traditionally, recombinant inbred lines are generated from only two parental generations. In a more recent project, recombinant inbred lines were derived from a genetically diverse set of eight founder inbred mouse lines [2, 9]. This experimental population was particularly designed to mimic the genetic diversity of humans, while keeping the advantages of a controlled population.

2.2.2 The Basics of QTL Mapping

A **quantitative trait locus (QTL)** is a location on the genome which hosts a gene that influences a certain quantitative trait. The major goal of QTL mapping is to identify such regions by means of statistical analysis. Considering an individual from a population, its trait value Y is a random variable which depends on the genetic background, as well as on many environmental factors. The **broad heritability** of a trait is defined to be the proportion of the trait's variance which can be explained by genetic factors.

To explain the basic principles of QTL mapping, we start with an extremely simple scenario. Consider a backcross design, and assume that there is exactly one QTL which influences the trait. Let $\mu_1 = E(Y|QTL = AA)$ denote the mean value of the trait when the QTL genotype is AA . Analogously, let $\mu_2 = E(Y|QTL = Aa)$. The coefficient $\beta = \mu_2 - \mu_1$ is called the effect size of the QTL. This is simply the expected increase in the trait value when allele a is substituted by A . Now consider a marker M which lies on the same chromosome as the QTL at a distance such that the probability of recombination between M and the QTL equals r . According to the

law of total probability, the expected value of the trait given the marker genotype is given by

$$E(Y|M = AA) = \mu_1(1 - r) + \mu_2r, \quad (2.5)$$

$$E(Y|M = Aa) = \mu_1r + \mu_2(1 - r). \quad (2.6)$$

Thus

$$E(Y|M = Aa) - E(Y|M = AA) = (\mu_2 - \mu_1)(1 - 2r)$$

and the difference between the mean trait values for individuals with different marker genotypes is different from zero as long as $r < 1/2$ (which means that the marker and the QTL are linked). Clearly, conditional on the marker genotype, the effect size is larger when the marker is closer to the QTL. This enables the detection of a QTL by identifying markers whose genotypes are associated with the trait. The details of statistical tests which can be used for this purpose will be discussed in Chap. 4.

As a second example, consider an intercross population with exactly one QTL. We define

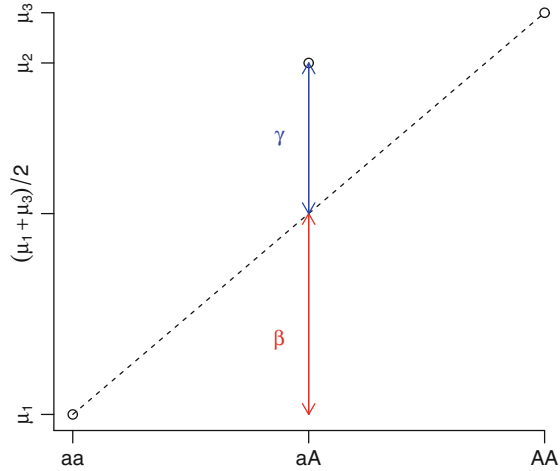
$$\mu_1 = E(Y|QTL = AA), \mu_2 = E(Y|QTL = Aa), \text{ and } \mu_3 = E(Y|QTL = aa).$$

The QTL is said to have a purely **additive effect** if $\mu_2 = (\mu_1 + \mu_3)/2$, which means that the average difference in the values of traits between individuals with genotypes AA and those with Aa is exactly the same as the average difference in the values of traits between individuals with genotypes Aa and those with aa . Otherwise, the QTL is said to have a **dominance effect**, defined by $\gamma = \mu_2 - (\mu_1 + \mu_3)/2$. If $\mu_1 \neq \mu_2 = \mu_3$, then the allele a is said to be dominant with respect to A . On the other hand, if $\mu_1 = \mu_2 \neq \mu_3$, then the allele a is called recessive. In general, the additive effect of a QTL is defined by $\beta = (\mu_3 - \mu_1)/2$. This corresponds to the coefficient of X in a regression model for the value of the trait and γ corresponds to the coefficient of Z . A graphical representation of additive and dominance effects is presented in Fig. 2.6. It is also possible that $\mu_2 \geq \max\{\mu_1, \mu_3\}$ or $\mu_2 \leq \min\{\mu_1, \mu_3\}$, which is referred to as *overdominance*.

In experimental populations there exists a very strong association between neighboring loci, as can be seen, for example, in Fig. 2.5. Therefore, to detect causal genes, it is usually enough to use approximately 10 markers on each chromosome. Based on the association studies discussed in the next sections, this enables us to minimize the problems resulting from multiple testing (see Sect. 3.1) and increase the power to detect QTLs. However, the strong association between neighboring markers also results in the rather low precision of estimators of the location of QTLs in experimental lines.

When compared to association studies, it is important to note that experimental populations give a researcher control over the genetic composition of the population. Usually, it is also much easier to control environmental influences on experimental populations than it is in natural populations, which is crucial for association studies.

Fig. 2.6 Graphical representation of the additive effect β and the dominance effect γ



2.2.3 Association Studies

The most important insight gained from the previous section is that a QTL can be detected if it is strongly linked to a genetic marker. This is a fundamental feature, which enables us to detect QTLs in experimental populations, where the experimental design gives us precise knowledge regarding the correlation between genetic loci on the same chromosome. The general idea of association studies is very similar: genetic markers strongly linked to influential genetic loci are expected to also be associated with the phenotype in question.

However, in contrast to QTL mapping, association studies are often performed with outbred populations, which means that the study sample does not include any close relatives. For outbred populations, the correlation structure between genetic loci is much more complicated than for experimental populations. There are no simple formulas corresponding to Eq. (2.3) which express the correlation between genetic loci as a simple function of the genetic distance. Instead, one analyzes **linkage disequilibrium** (LD), which is a measure of the nonrandom association between different genetic loci. For an overview of measures of LD see [3].

Two markers in LD cannot be treated as independent variables, since they are correlated. The somewhat complex theory underlying LD in outbred populations is a subject of the theory of population genetics. The most relevant source of LD in association studies is the genetic linkage between markers located close to each other on the same chromosome. As in experimental populations, a genetic marker can be associated with a trait, although it does not directly affect that trait. It is sufficient that such a marker is closely linked to a QTL. However, the local structure of LD tends to be rather complicated, as can be seen in Fig. 2.7, which is a heatmap based on data from the HapMap project. The color of this heatmap represents the LD measure R^2 , which is simply defined to be the square of the Pearson correlation

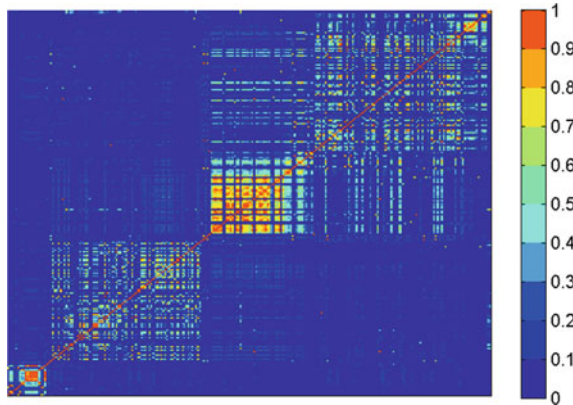


Fig. 2.7 Heatmap illustrating the LD pattern for 90 individuals from the CEU HapMap population (of Central European ancestry). The color code corresponds to the LD measure R^2 between the first 250 adjacent SNPs of the ENCODE region ENM010 (after removing duplicate sequences)

coefficient (see Sect. 7.4) between the genotypes of SNPs (coded as $X \in \{-1, 0, 1\}$ as in Sect. 2.2.1.2).

The HapMap project has been crucial in developing a better understanding of LD patterns within the human genome [13, 14]. In the second phase of the project, genotypes were obtained from 269 individuals from four populations: 90 Yoruba from Ibadan, Nigeria (YRI); 90 U.S. residents of Central European ancestry (CEU); 45 Japanese from Tokyo (JPT); and 44 Chinese from Beijing (CHB). The aim was to obtain a comprehensive map of the human genome. After completing phase II of the project in 2007, over 3.1 million SNPs had been genotyped. The ENCODE project [12] considered ten regions in particular, for which we have highly accurate genetic maps. Within these regions of approximately 500 kilo base pairs, almost all the SNPs known at the time were genotyped within this study group.

Other popular measures of pairwise LD in the population genetics literature are often based not on genotype information, but require additional knowledge about **phased haplotypes**, i.e., one has to know which copy of the parental chromosome each marker allele belongs to.

Let us consider two genetic loci, A and B , and the corresponding dummy variables Y_{Ai} and Y_{Bi} , which are equal to 1 if the allele at the i th phased haplotype at the corresponding locus is the reference one and 0 otherwise (e.g., $i = 1$ may correspond to the maternal haplotype and $i = 2$ to the paternal haplotype). Then the frequencies of the reference alleles at these loci are given by $p_A = \bar{Y}_A$ and $p_B = \bar{Y}_B$, where \bar{Y}_A and \bar{Y}_B denote the average of Y_{Ai} and Y_{Bi} over all $2n$ phased haplotypes (where n is the number of diploid individuals). Moreover, the percentage of haplotypes for which the reference allele appears at both locations is given by the average of the products of Y_{Ai} and Y_{Bi} ; $p_{AB} = \overline{Y_A Y_B}$. Two classical measures of LD between A and B , Lewontine's

D' and r^2 , are based on the statistic $D = \text{Cov}(Y_A, Y_B) = p_A p_B - p_{AB}$. According to Lewontine's D' , D is scaled according to

$$D_{\max} = \begin{cases} (1 - p_A)(1 - p_B) & \text{if } D < 0 \\ \min\{p_A(1 - p_B), p_B(1 - p_A)\} & \text{if } D > 0, \end{cases} \quad (2.7)$$

i.e., $D' = D/D_{\max}$. To derive r^2 , the square of D is divided by the product of the variances of Y_A and Y_B ; $r^2 = \frac{D^2}{p_A(1-p_A)p_B(1-p_B)}$. Thus r^2 is simply the square of the standard Pearson correlation coefficient between Y_A and Y_B . Now, observe that the variable X_A describing the genotype at locus A can be represented as $X_A = Y_{A1} + Y_{A2} - 1$, where Y_{A1} and Y_{A2} are the dummy variables corresponding to the haplotypes. It is easy to check that when Y_{A1} and Y_{A2} are independent, then the population correlation coefficient between X_A and X_B equals the correlation coefficient between Y_A and Y_B . Therefore, in real life situations the correlation coefficient R^2 between the genotypes at two loci is typically very close to r^2 .

Figure 2.7 nicely illustrates the rather complex local LD patterns in humans. There appears to be some kind of block structure, where a number of neighboring SNPs tend to be all highly correlated with each other. However, this pattern of correlation is not necessarily completely in accordance with physical distance, which has already been pointed out in [13]. For example, the large LD block in the middle is pervaded by several blue stripes. This indicates that there is a set of highly correlated SNPs, but in between there is a number of other SNPs which are not at all correlated with that set.

Such models of haplotype blocks provide a useful first approach to looking at LD structure in outbred populations. However, in reality the situation appears to be even more complicated [14]. In any case, the mere existence of local LD enables us to adapt the idea of association studies to outbred populations. Compared with the experimental populations discussed previously, linkage is only observed at a much smaller scale of genetic distance. As a result, a much larger number of genetic markers are necessary to perform association studies. On the other hand, association studies enable much greater precision in localizing influential genes, precisely because LD does not extend over larger regions of the genome.

Association studies have a long history within genetic research, although linkage studies based on family pedigrees used to be more prominent in the past. While linkage studies have been quite successful for locating the QTLs of traits that are controlled by a single locus (Mendelian traits), they have turned out to be far less successful for locating QTLs for traits that are determined in a more complex manner. In the twentieth century, association studies were often rather limited by the fact that an insufficient number of genetic markers were available. In many cases, only a relatively small number of candidate polymorphisms, which were suspected in advance to be related to some trait, could be analyzed. In **candidate gene studies**, researchers are only interested in polymorphisms lying within the region of genes suspected to affect the trait in question. Preselection of the candidate genes is often based on a biological understanding of their functioning. On the other hand, there

might exist some knowledge stemming from previous studies about genetic regions associated with a trait. In such cases, a follow up association study might run under the name of “fine mapping.”

Today, a sufficiently dense map of genetic markers (usually SNPs) is available to carry out **genome wide association studies (GWAS)** among many species. In particular, we have already mentioned the HapMap project, based on the human population, which has mapped several million SNPs across the whole genome. More recently, the number of known SNPs has been further increased by the 1000 genomes project [17], which aims at sequencing the whole genome of 2500 individuals from about 25 populations around the world. Knowledge about such polymorphisms is important when designing a genetic map, but equally important is the question of how to determine the genetic variants of individuals participating in a study. Within the last two decades, microarray and sequencing technology have made rapid progress and brought down the costs of determining individuals’ genotypes. In Chap. 5 we will discuss the technological aspects underlying GWAS in more detail, in particular, the microarray technology which enables determining millions of genetic variants in one experiment.

An interesting question in GWAS is whether it is really necessary to work with all known polymorphisms. Although the most recent technology allows us to genotype more than 2 million SNPs, this is still less than 20 % of the variants available today. On the other hand, it is well known that in regions of strong LD, a small number of phased haplotypes (so-called common haplotypes) comprise a very large majority of all haplotypes (see for example [19]). When designing an association study, one has to decide which SNPs one should genotype in patients. Similarly, when designing a SNP array, one has to decide which SNPs to put on that array. Clearly, it is not advisable to consider SNPs which are highly correlated, as they will each provide almost the same information. This leads to the idea of **tag SNP selection**: starting from an extensive set of SNPs, one looks for a minimal subset of SNPs, so-called tag SNPs, which contain as much information as possible. A large number of algorithms for tag SNP selection are available (see [6] for a brief review). A set of tag SNPs covering the whole genome can then be used to create SNP arrays.

In view of the latest developments in next generation sequencing, we can look ahead toward association studies where the complete genetic information regarding individuals is available. In theory, the question of LD would then be resolved, because all genetic variation (at the level of DNA) would be known. However, apart from the fact that occasional errors in sequencing lead to imperfect information, many other difficulties presently affecting GWAS will remain. In particular, the problem of multiple testing will become even worse. In Chap. 3, the statistical theory of multiple testing will be comprehensively discussed. Here, we only want to mention that due to the tremendous number of SNPs (or other types of genetic variation) considered in association studies, it is very likely that one observes an association between a trait and a marker just by chance. The only remedy against this intrinsic statistical problem is to perform very large scale studies, and it has become more or less standard in GWAS to consider study groups with several thousand participants.

2.2.3.1 Design Questions in Association Studies

So far, we have been mainly concerned with the basic ideas underlying population association studies. Ideally, these are based on a large number of unrelated individuals. Unrelated essentially means that relationships are distant enough so that no linkage due to the relatedness of individuals is observed. To guarantee that this is the case, it is advisable to perform statistical tests which rule out unknown family relationships between each pair of participants. Otherwise, the theoretical properties of the statistical tests applied might become distorted. For example, undetected relationships between participants of a study could increase the number of false positives.

Association studies can be performed based on quantitative traits (as in QTL mapping), but more often one deals with dichotomous traits, usually characterizing an individual's status with respect to a certain disease. In so-called **case control studies**, one considers samples of affected cases and unaffected controls. Often, it is relatively easy to recruit cases for association studies, but it might be more difficult to find controls who are prepared to be genotyped. Also, in view of financial restrictions, it is often easier to use a control group from the general population, for whom genomic data should be available in a reference database. However, using such a **case random design**, the presence or absence of the disease in question in members of the “control” group has not been ascertained. However, if the prevalence of the disease is small, then there should be hardly any difference between the effectiveness of case control design and case random design.

One important assumption underlying association studies is that the study population is **panmictic**, which means that all individuals of the opposite sex in the population are potential partners. In practice, it is often the case that random mating cannot be assumed. Within the study population there might be relatively homogeneous subgroups, for example, ethnic subgroups, social subclasses or geographically separated groups. As we will see in Sect. 5.2.3, the resulting **population structure** can have a serious effect on the statistical analysis of association studies if not accounted for appropriately.

2.2.4 Other Types of Study

This book will mainly focus on the statistical analysis of experimental populations (Chap. 4) and on panmictic populations (Chap. 5). One reason for this choice is that we will emphasize a particular approach to model selection, which has been fully developed for these two types of study in terms of both statistical theory and software. There exist a number of additional types of study. In this section, we will briefly discuss admixture mapping and some aspects of data from families. Data from families have been used extensively for linkage analysis, but more recently there has also been some interest in association studies based on such data. In this case, it

seems possible to extend the statistical methods described in Chaps. 4 and 5, but the details still have to be worked out in future research.

2.2.4.1 Admixture Mapping

We mentioned above that population structures can lead to problems in association studies. However, in certain situations it can be the basis for study designs with desirable qualities. In recent years, the analysis of admixture populations has gained popularity. Such populations result from previously separated subpopulations becoming mixed. African-Americans and Latinos are perhaps the most prominent example of a pair of such subpopulations, where gene flow between these populations started only a few hundred years ago.

In admixture mapping, it is assumed that only the local LD typically found in outbred populations exists within the original subpopulations (see Fig. 2.7). From a genetical point of view, mating between two outbred populations has some similarities to the experimental crosses of inbred populations discussed in Sect. 2.2.1. Due to the process of repeated crossovers over several generations, an individual from an admixed population will have relatively long strands of DNA, where different strands stem from different ancestral populations and the length of these strands will depend on the mixing history.

The idea of admixture mapping is to use ancestral information to localize regions which influence a trait. Assume there are two founding subpopulations A and B . Then at a certain genetic locus there are three possible states of ancestry, AA , AB or BB . If a certain risk allele occurs in the ancestral population A more frequently than in B , then one would expect that the ancestral state A is observed in affected individuals from the admixed population more often at the location of that risk allele compared to other genomic regions.

One potential advantage of admixture mapping is that within recently admixed populations these strands from ancestral populations are still rather long. Hence, compared with association studies, a substantially smaller number of genetic markers are needed. Also, admixture mapping is able to locate influential regions, even when studying only cases without a control group (e.g., see [8]). On the other hand, admixture mapping only works for risk alleles which have substantially different frequencies in the ancestral populations. We recommend [18] as an introduction to such an approach and [20] to learn more about the statistical methods involved in admixture mapping.

2.2.4.2 Data from Families

As mentioned above, association studies are commonly performed based on pan-mictic populations and aim to find a correlation between a genetic marker and the trait in question. Data from families are normally analyzed using linkage analysis, which is based on slightly more indirect logic. In linkage studies, one tries

to identify loci which cosegregate with the trait (are inherited along with the trait) within families. For a given pedigree of a family, one can compute the joint probability of specific marker genotypes and disease status. Based on such computations, one can test whether a genetic marker is in LD with a genetic locus which directly affects the trait.

An important concept in linkage analysis is being **identical by descent** (IBD). Among relatives, alleles are IBD if they arose from the same allele of a common ancestor. Tests in linkage analysis are often based on the IBD configuration, but it is not always possible to determine the IBD status of all individuals in a pedigree. This results from the problem of phasing, which is usually solved using maximum likelihood methods. However, the necessary computations become rather involved for complex pedigrees. In view of this, designs like the **affected sib-pair** method are rather popular. This is based on comparing the similarity of siblings who share the same biological mother and father. A more detailed introduction to the mathematical and statistical aspects of analyzing pedigree data can be found, for example, in [10, 15] or [16].

References

1. Broman, K.W.: The genomes of recombinant inbred lines. *Genetics* **169**, 1133–1146 (2005)
2. Chesler, E.J., Miller, D.R., Branstetter, L.R., Galloway, L.D., Jackson, B.L., Philip, V.M., Voy, B.H., Culiati, C.T., Threadgill, D.W., Williams, R.W., Churchill, G.A., Johnson, D.K., Manly, K.F.: The collaborative cross at Oak Ridge National Laboratory: developing a powerful resource for systems genetics. *Mamm. Genome* **19**, 382–389 (2008)
3. Devlin, B., Risch, N.: A comparison of linkage disequilibrium measures for fine-scale mapping. *Genomics* **29**, 311–322 (1995)
4. Feingold, E.: Methods for linkage analysis of quantitative trait loci in humans. *Theor. Popul. Biol.* **60**, 167–180 (2001)
5. Fisher, R.A.: *The theory of inbreeding*, 2nd edn. Academic Press, New York (1965)
6. Frommlet, F.: Tag SNP selection based on clustering according to dominant sets found using replicator dynamics. *Adv. Data Anal. Classif.* **4**, 65–83 (2010)
7. Haldane, J.B.S., Waddington, C.H.: Inbreeding and linkage. *Genetics* **16**, 357–374 (1931)
8. Hoggart, C.J., Shriver, M.D., Kittles, R.A., Clayton, D.G., McKeigue, P.M.: Design and analysis of admixture mapping studies. *Am. J. Hum. Genet.* **274**(5), 965–978 (2004)
9. Iraqi, F.A., Churchill, G., Mott, R.: The Collaborative Cross, developing a resource for mammalian systems genetics: a status report of the Wellcome Trust Cohort. *Mamm. Genome* **19**, 379–381 (2008)
10. Lange, K.: *Mathematical and Statistical Methods for Genetic Analysis*. Springer (1997)
11. Scheet, P., Stephens, M.: A fast and flexible statistical model for large-scale population genotype data: applications to inferring missing genotypes and haplotypic phase. *Am. J. Hum. Genet.* **78**, 629–644 (2006)
12. The ENCODE Project Consortium: the ENCODE (ENCyclopedia of DNA Elements) Project. *Science* **306**, 636–640 (2004)
13. The International Hapmap Consortium: a haplotype map of the human genome. *Nature* **437**, 1299–1320 (2005)
14. The International Hapmap Consortium: a second generation human haplotype map of over 3.1 million SNPs. *Nature* **449**, 851–862 (2007)

15. Thompson, E.A.: Pedigree Analysis in Human Genetics. Johns Hopkins University Press, Baltimore (1986)
16. Thompson, E.A.: Statistical Inferences from Genetic Data on Pedigrees. In: The Proceedings of the NSF-CBMS regional conference series in probability and statistics, vol. 6. I MS, Beachwood, OH (2000)
17. Via, M., Gignoux, C., Burchard, E.G.: The 1000 genomes project: new opportunities for research and social challenges. *Genome Med.* **2**, 3 (2010)
18. Winkler, C.A., Nelson, G.W., Smith, M.W.: Admixture mapping comes of age. *Annu. Rev. Genomics Hum. Genet.* **11**, 65–89 (2010)
19. Zhang, K., Sun, F.: Assessing the power of tag SNPs in the mapping of quantitative trait loci (QTL) with extremal and random samples. *BMC Genet.* **6** (2005)
20. Zhu, X., Tang, H., Risch, N.: Admixture mapping and the role of population structure for localizing disease genes. *Adv. Genet.* **60**, 547–569 (2008)

Phenotypes and Genotypes

The Search for Influential Genes

Frommlet, F.; Bogdan, M.; Ramsey, D.

2016, XII, 222 p. 33 illus., 20 illus. in color., Hardcover

ISBN: 978-1-4471-5309-2