

Introduction to Experimental Genetics

Variability in observed phenotypes may result from a variety of factors – inherited as well as environmental. The blend of all these influences gives rise to the unique being every living creature is. Still, the main role of science is to identify the rules which unite seemingly unrelated phenomena. The role of genetics is no different. Its first, and most important, task is to identify the major factors that give rise to different phenotypical characteristics. Once these major factors have been identified, the investigation can be carried on in order to identify genes having secondary effects.

The basic strategy in experimental sciences is to isolate the phenomena being investigated and study them under controlled conditions. Ideally, an experiment will involve measurement of the phenomena, taken at various levels of one or a small number of potentially influencing factors with the levels of other factors kept fixed. Thence, observed variation in the phenomena can be attributed to the variation in that single explanatory factor. Experimental genetics applies this approach. Effort is made to ensure uniform environmental conditions, so one hopes that environmental effects on the phenotype are minimized. Moreover, the use of an experimental cross of inbred strains (see below) potentially reduces the variability in the genetic background, and thus facilitates the dissection of the genetic factors that give rise to the observed phenotypic variation. (This simplification brings with it the disadvantage that there may well be genetic variability in nature that is not represented in a given set of inbred strains.)

In reality, the ideal experiment is almost never feasible. Environmental condition may vary slightly from one stage of the experiment to the next. Pure strains may be contaminated. Some factors may not be possible to isolate from others, thus forcing the investigation of several factors jointly. And on top of that, measurement errors may introduce unwanted error into the system. Consequently, precautionary measures need to be taken. These may include the use of experimental controls, taking repeated measurements, and reproducing the results in other laboratories. The analysis of the outcomes

of the experiments must take into account the potential effect of unplanned factors. This is the role of statistical models and statistical inference.

Models from statistical genetics are used in order to incorporate under one roof the investigated effects of the genetic factors, measurement errors, and a whole range of influencing factors that may have not been taken into account during the planning of the experiment. The analysis of the outcomes of the experiment, and the conclusions, are based on applying the principles of statistical inference to these models. The realization that eventually the outcomes of the experiment will be investigated with statistical tools should have an effect on the way the experiment is planned. A thoughtful design of an experiment may increase its statistical efficiency. A poor design may be wasteful, or may not produce enough information to attain its stated goal.

Our discussion concentrates on mouse genetics, but the methods are valid more generally. The mouse model was selected due to its importance in relation to human genetics. The rest of this chapter provides some background in specific statistical models and experimental designs for genetic mapping based on crosses of inbred strains. Because we are not able to do breeding experiments with humans, human genetics involves more complex statistical considerations that will be investigated in later chapters. The first section below outlines some basic facts regarding the mouse model and its genetics. The material in this section is mainly borrowed from the excellent textbook by Lee Silver [77]. (See the link <http://www.informatics.jax.org/silver/> for an online version of the book.) The second section deals with a statistical model of the relation between genetic factors and the observed phenotype. Like reality, statistical models can become quite complex. However, even simple models can provide insight and lead to useful analysis. The model that will be discussed here is as simple as such models can get. Thus, it is perfect for the introduction of the basic concepts. Yet, even in its simplest form, the model is frequently used by researchers – a testimony to its value. Some simulations are conducted in accordance with the model in order to demonstrate the effect of genetic factors on the distribution of the phenotype. In the third section some common experimental designs in mice are described. The merits and drawbacks of the different designs will be revisited in later chapters after going into the details of the statistical tools and their properties. Finally, a short and non-technical description of the genetic markers that are commonly used today is provided.

2.1 The Mouse Model

Already in the early days of modern genetics, the house mouse was identified as a perfect model for genetic investigation in mammals. The house mouse is a small, easy to handle animal. It is relatively inexpensive to maintain, it breeds easily, and it has a high rate of reproduction. A significant portion of biological research is aimed at understanding ourselves as human beings. Although many

features of human biology at the cell and molecular levels are shared across all living things, the more advanced behavioral and other characteristics of human beings are shared only in a limited fashion with other species or are unique to humans. In this vein, the importance of mice in genetic studies was first recognized in the intertwined biomedical fields of immunology and cancer research, for which a mammalian model was essential. Today, specially developed mouse strains serve as models for many human traits, e.g., obesity or diabetes.

The mouse genome, like the genomes of most mammals, contains about three billion base-pairs (bp). It is organized in 19 pairs of homologous autosomal chromosomes, compared to the 22 pairs in human, and a single pair of sex chromosomes. Yet, there is a large amount of homology between the mouse and the human genes. Large chunks of DNA, 10–20 million bp in length, remained intact during evolution, and are practically identical in both species. As a matter of fact, the entire human genome can be, more or less, recovered by cutting the mouse genome into 130–170 pieces and reassembling them in a different order.

At the genetic level, processes of meioses, mating, and reproduction in the two species are similar. In particular, in mice like in human, autosomal chromosomes may experience crossovers during meiosis. This process of recombination mixes up the genetic material that is passed on from the parent to the offspring. Owing to recombination, the genetic contribution of the parent is a random mosaic of segments originating from the two grandparents. However, the rate of crossovers per base pair during meiosis in mouse is about half the rate in human.

The founding father of mouse research was W. E. Castle, who (intentionally) brought the mouse into his laboratory at the beginning of the 20th century. Subsequently, he, and his many students, began developing genetically homogeneous inbred lines of mice. These pure inbred lines became a very valuable resource and the key to the success of the mouse model in genetics. Genetically homogeneous strains provide the means to control for the effect of genetic factors. Moreover, since all mice of the same line are genetically identical, results of experiments carried out in a specific laboratory can be compared to results of experiments from other laboratories. A major source of such genetically pure strains is the Jackson Laboratory in Bar Harbor, Maine. This laboratory was founded in 1929 by C. Little, a student of Castle's, with the aim of promoting mouse research; it serves to this day as a center of mouse research.

Genetically homogeneous inbred strains are created by a process of successive brother–sister mating. Random drift in finite inbred populations eventually results in the fixation of a given locus, namely in the extinction of all other alleles. Once a locus is not polymorphic, it remains so in all subsequent generations. With additional brother–sister mating, the genomes in the population become less and less polymorphic. The formal definition of an inbred strain requires at least 20 generations of strict brother–sister mating. Some

of the classical inbred strains have a history of more than 100 generations of inbreeding. See Chap. 3 for a more systematic discussion of inbreeding.

2.1.1 A Quantitative Trait Locus (QTL)

We consider measurement of a continuous trait, such as body weight or blood pressure, in a population of mice. Denote the level of the measurement for a randomly selected mouse by y . One usually observes that such measurements show variability across the population. Some of this variability may be attributed to genetic factors. The task is to model the overall genetic contribution, and the genetic contribution of each specific locus, to the overall variance.

We assume that the phenotypic value y is a simple summation of the mouse's genotype and environment influences:

$$y = m + G + E, \quad (2.1)$$

where m is a constant, G denotes the effect of all genes contributing to the phenotype, and E denotes the environmental effects. By writing m separately we can assume without loss of generality that both G and E have mean value 0, so m is in fact the average phenotypic value. We also make the critical assumption that G and E are uncorrelated. The variance of G , say σ_G^2 , is called the genetic variance; the variance of y , σ_y^2 is the phenotypic variance. From our assumption that G and E are uncorrelated, it follows that σ_y^2 is the sum of σ_G^2 and the variance of E , σ_E^2 . An important quantity is the heritability H^2 defined to be the ratio σ_G^2/σ_y^2 of the genotypic to phenotypic variance. It measures the percentage of variation in the phenotype that has a genetic origin.

We now want to make more specific assumptions about the contribution to G arising from one genetic locus. Consider a given polymorphic genetic locus, specifically, a bi-allelic locus with its two alleles denoted by A and a . The genotype at the given locus is one of the three types: AA , Aa , or aa . Consequently, the underlying population can also be subdivided into three subclasses according to these three types. The model we propose may assign different average measurements for each of these subclasses. However, the variance of the measurement is assumed to be unchanged across the three subclasses. Let x_M be the number of A alleles (0 or 1) inherited from the mother and x_F the number inherited from the father. Observe that $x = x_M + x_F$, the total number of A alleles in the genotype, can take on the values 0, 1, or 2. Consider a model of the form

$$y = \mu + \alpha(x_M + x_F) + \delta|x_M - x_F| + e, \quad (2.2)$$

where e is a zero mean random deviate. Note that $|x_M - x_F| = 1$ if, and only if, $x = 1$, namely that the mouse is *heterozygous*. The term μ is the mean level of y for an *aa-homozygote*. The mean level for an *AA-homozygote* is $\mu + 2\alpha$,

and the mean level for an heterozygote is $\mu + \alpha + \delta$. The locus is said to be *additive* if $\delta = 0$, since in this case each A allele adds the amount α to the average phenotype. If $\delta = \alpha$, the allele A is said to be *dominant*; it is called *recessive* if $\alpha = -\delta$. These terms are consistent with the usage for qualitative traits. For a dominant locus, a single A allele produces the full genetic effect; for a recessive locus a single A allele produces no effect, while two A alleles do produce an effect. Note that A is dominant if and only if a is recessive.

The term $x = x_M + x_F$ is the number of A alleles in a randomly selected mouse. Both x_M and x_F are Bernoulli random variables. We assume here that they are independent with the same probability p to take the value 1 (indicating an A allele). Then the distribution of x is binomial, namely $\Pr(x = 2) = p^2$, $\Pr(x = 1) = 2p(1 - p)$, and $\Pr(x = 0) = (1 - p)^2$, where p is the frequency of allele A in the population. It follows that the mean value of x is $2p^2 + 2p(1 - p) + 0 = 2p$. Therefore, the overall mean of the phenotype is $m = \mu + 2p\alpha + 2p(1 - p)\delta$. (The assumption that x_M and x_F are independent Bernoulli variables can be derived from the notion of *Hardy-Weinberg Equilibrium*, which will be discussed in the next chapter.)

The residual e incorporates all remaining factors that contribute to the variability. Such factors can include the genetic contribution from loci other than the one we investigate, as well as environmental factors. We assume that e is *uncorrelated* with the other terms on the right-hand side of (2.2). This assumption may not be satisfied. In most cases it rules out the possibility of other genes on the same chromosome that contribute to the trait. Such genes would be *linked* to and hence correlated with the investigated locus. (See the discussion of recombination in Chap. 3.) In the special case where the locus explicitly modeled in (2.2) is the *only* genetic locus contributing to the trait, then e in (2.2) is the same as E in (2.1); and since G has mean 0, it can be written explicitly as $G = \alpha(x - 2p) + \delta[x_M - x_F - 2p(1 - p)]$.

Often in what follows we assume also that e is normally distributed, although this assumption is not strictly necessary. Because of the central limit theorem, it would be a reasonable assumption if e is made up of a sum of approximately independent small effects, either genetic or environmental. It would not be satisfied, however, if there is another major gene having a substantial effect in addition to the one modeled explicitly in (2.2). We will return to this point below.

By simple, but somewhat tedious algebra (see Prob. 2.4 below) one can rewrite the model (2.2) in the form:

$$y = m + \{\alpha + (1 - 2p)\delta\} \times [(x_M - p) + (x_F - p)] - \{2\delta\} \times [(x_M - p)(x_F - p)] + e, \quad (2.3)$$

and show that $[(x_M - p) + (x_F - p)]$ and $[(x_M - p)(x_F - p)]$ are uncorrelated by virtue of the assumption that x_M and x_F are independent. Since by assumption e is also uncorrelated with these terms, the variance of y can be written as the sum of three terms:

$$\sigma_y^2 = \sigma_A^2 + \sigma_D^2 + \sigma_e^2, \quad (2.4)$$

where

$$\sigma_A^2 = 2p(1-p)[\alpha + (1-2p)\delta]^2, \quad \sigma_D^2 = 4p^2(1-p)^2\delta^2, \quad \sigma_e^2 = \text{var}(e). \quad (2.5)$$

The term σ_A^2 is called the locus specific *additive variance*, while σ_D^2 is called the locus specific *dominance variance*. Note the potential confusion with the terminology introduced above. The allele A is additive if and only if the dominance variance is 0. If A is dominant *or* recessive, the dominance variance is positive. In the very important special case of an *intercross*, defined below in Sect. 2.2, $p = 1/2$, so $\sigma_A^2 = \alpha^2/2$, and $\sigma_D^2 = \delta^2/4$.

A measure of the importance of the locus under study is the *locus specific heritability*, denoted by h^2 and defined to be the ratio $(\sigma_A^2 + \sigma_D^2)/\sigma_y^2$. In the special case that only a single gene contributes to the trait $h^2 = H^2$, the heritability of the trait.

2.1.2 Simulation of Phenotypes

Let us explore the effect of the different model parameters on the distribution of the phenotype in a population. For the exploration we use R, which was introduced in the previous chapter. Start with formation of the vector of mean values, for the case where $\mu = 5$, $\alpha = 1$, and $\delta = -1$ (i.e., a recessive model). For each animal the mean expression level is equal either to 5 or to 7, depending on the animal's genotype:

```
> n <- 6; p <- 0.5; x <- rbinom(n,2,p)
> mu <- 5; alpha <- 1; delta <- -1
> mu + alpha*x + delta*(x==1)
[1] 7 5 5 7 5 5
```

A normal residual term is added to the mean:

```
> sig <- 0.5
> mu + alpha*x + delta*(x==1) + sig*rnorm(n)
[1] 8.042733 5.351743 4.859863 7.133684 4.671937 5.558227
```

The default application of the function “**rnorm**” generates independent standard normal random variables. Multiplying by the standard deviation and adding the mean transforms the distribution to having the given mean and standard deviation. An alternative way of obtaining the same distribution is by using the “**mean**” and “**sd**” parameters of the function:

```
> rnorm(n, mu + alpha*x + delta*(x==1), sig)
[1] 8.312484 5.351396 4.635130 6.936316 4.829702 5.295540
```

Yet, a third way would be to have:

```
> mu + alpha*x + delta*(x==1) + rnorm(n, sd=sig)
[1] 6.982847 6.354027 5.177322 6.852138 5.111498 4.644922
```

Observe that we must introduce the standard deviation with the “`par_name = par_value`” argument assignment format since the parameter “`sd`” is not the second parameter of the function “`rnorm`”.

Let us put some real action into the story by simulating a population of 100,000 animals according to the given genetic model, under two different scenarios for non-genetic variability:

```
> n <- 10^5; p <- 0.5; x <- rbinom(n,2,p)
> mu <- 5; alpha <- 1; delta <- -1; sig <- 0.5
> y <- mu + alpha*x + delta*(x==1) + sig*rnorm(n)
> h2 <- var(alpha*x + delta*(x==1))/var(y)
> plot(density(y),main=
+   paste("A recessive model:\n h^2 = ",round(h2,3),sep=""))
```

This produces the density function on the left in Fig. 2.1. The density on the right is produced by:

```
> mu <- 5; alpha <- 1; delta <- -1; sig <- 1
> y <- mu + alpha*x + delta*(x==1) + sig*rnorm(n)
> h2 <- var(alpha*x + delta*(x==1))/var(y)
> plot(density(y),main=
+   paste("A recessive model:\n h^2 = ",round(h2,3),sep=""))
```

The object “`h2`” stores the value of h^2 , i.e., the fraction of the genetic variance for the modeled gene within the total variance. The function “`density`” produces an estimate of the density of the population, based on the values stored in the vector `y`. The resulting object is then plotted with the function “`plot`”.

The title for the figure is set with the argument “`main`” of the function “`plot`”. The assignment of this argument is a character string. Character strings are entered using either double (") or single (') quotes. The function “`paste`” takes an arbitrary number of arguments and concatenates them one by one into character strings, with a separating string determined by the argument “`sep`”. Numbers are converted into character strings. The symbol “`\n`” enters a line break.

The distributions of the phenotypes in both populations are given in Fig. 2.1. Both pictures represent a mixture of two normal distributions. The mean of each of the normal distributions that form the mixture is the same (5 and 7), and the fraction is the same (3/4 and 1/4). The only difference between the two cases is the magnitude of the variance attributed to non-genetic factors (1/4 versus 1). Note that this small change is enough to change a bi-modal distribution into a distribution with a single mode. In fact, a casual glance at the second density function in Fig. 2.1 suggests that it is roughly normal, although a more careful inspection shows that it is slightly skewed to the right.

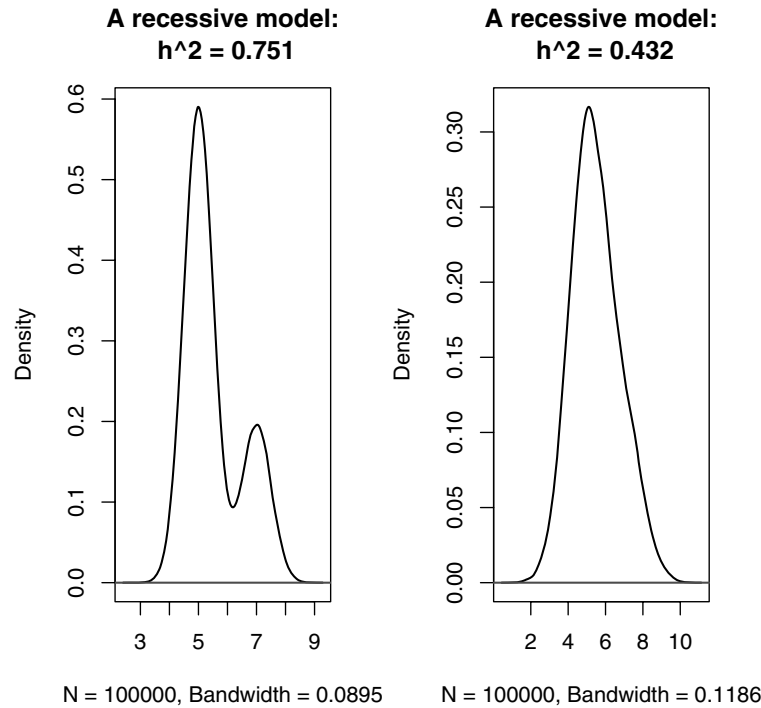


Fig. 2.1. Distributions of QTL phenotypes

2.2 Segregation of the Trait in Crosses

Crosses between inbred strains are the bread and butter of experimental genetics. Inbred strains are specially developed to be homozygous at all loci throughout the genome, in contrast to outbred populations which are often polymorphic. Standard approaches for dissecting an heritable trait in experimental genetics often involve crossing inbred strains. The genetic contribution to a trait cannot be demonstrated by looking at individuals from a single inbred strain alone, since in principle all members of the same strain are genetically identical and the observed phenotypal variability among them has to be attributed to environmental factors. Therefore, in order to map genetic factors one must select at least two different strains, which show different levels of expression of the phenotype under the same environmental conditions. In many cases one can screen the commercially available inbred strains for an appropriate pair of strains, which are phenotypically distinct. Careful crosses between the selected strains produce the population that is used for the genetic mapping.

The two most popular protocols for crossing inbred strains are the *backcross* and the *intercross*, which we describe below. However, before going into the details of the two different breeding schemes, let us introduce the standard terminology. Practically all breeding experiments start with an *outcross*, a mating between two animals or strains considered unrelated to each other. Specifically, we consider here an outcross between two distinct inbred strains that show a phenotypic difference. The resulting offspring are called the first filial generation, or F_1 . Consider any genetic locus where the two strains differ. Recall that inbred strains are homozygous at all loci. Say that the genotype of one of the inbred strains at a given locus is AA , and the genotype of the other inbred strain is aa . It follows that the genotype of the F_1 generation at the polymorphic site must be Aa , since each parent passes on one of its alleles to the offspring. Consequently, F_1 mice have a fixed genetic composition – all are heterozygous at all polymorphic loci.

A backcross is obtained by mating the F_1 offspring with mice from either one of the original inbred strains. Note that there are two possible types of backcross, depending on the choice of the inbred strain for the cross, denoted here by the $AA \times Aa$ backcross and the $aa \times Aa$ backcross. (As a matter of fact, one can further divide the backcross breeding scheme based on the sex of the F_1 mice in the cross. This may be important if a sex-linked trait or imprinting is considered. However, we ignore these possibilities in the sequel.)

Consider the $aa \times Aa$ backcross design in the context of the simple QTL model described above in (2.2). Assume the mother is inbred and the father is the F_1 . Then x_M is always equal zero, but x_F may be zero or one. The backcross offspring are either aa homozygous, which corresponds to $x_F = 0$, or Aa heterozygous, which corresponds to $x_F = 1$. The probability of each of the values is $1/2$, with corresponding phenotypic mean levels of μ and $\mu + \alpha + \delta$. The resulting regression model is given by

$$y = \mu + (\alpha + \delta)x_F + e, \quad (2.6)$$

with a similar equation for the $AA \times Aa$ backcross. The offspring are either AA homozygous or Aa heterozygous. In this case, the variable $1 - x_F$ has a Binomial(1, 0.5) distribution, and the regression model takes the form:

$$y = (\mu + 2\alpha) + (\delta - \alpha)(1 - x_F) + e. \quad (2.7)$$

Both models (2.6) and (2.7) have a similar form. However, their statistical properties may be quite different depending on the relations between α and δ . Since a Bernoulli random variable with $p = 1/2$ has variance $p(1 - p) = 1/4$, the variance component associated with the genetic factor is $(\alpha + \delta)^2/4$ for the first model and $(\alpha - \delta)^2/4$ for the second. Consequently, the locus specific heritability is larger for the first model if α and δ have the same sign, and vice versa if they have opposite signs. For the additive model ($\delta = 0$) the ratios are equal.

The intercross is a result of the mating of an F_1 male and an F_1 female. In terms of the notation, we will refer to the intercross as the $Aa \times Aa$ cross. The

term F_2 may also be used. (Subsequent generations of mating are denoted F_n , where n is the number of generations since the initial outcross.) The offspring of the $Aa \times Aa$ intercross can have any one of three genotypes. The distribution of the genotypes follows the ratios 1:2:1, thus x has a binomial distribution, $B(2, 1/2)$. It was shown above (Equation (2.5) and the following sentences) that the variance component associated with the given locus is $\alpha^2/2 + \delta^2/4$.

We turn now, with the aid of a small simulation study, to a demonstration of the segregation of the trait in a cross. The simulation will generate segregation of the alleles from parents to offspring and the resulting expression of the phenotype in the offspring. Recall that a parent carries two alleles at a given locus, one inherited from the grandfather and the other inherited from the grandmother. Only one of these two alleles will be passed on to the parent's offspring. According to Mendel's first law of segregation each of the alleles has an equal chance to be passed on. For example, if the parent is Aa at the given locus, then it will pass with equal probabilities either the allele A or the allele a . Of course, if the parent's genotype is AA (aa) it will pass on the allele A (respectively a) with certainty.

Assume the parent is an F_1 mouse:

```
> n <- 9
> pat <- rep("A",n)
> mat <- rep("a",n)
> pat
[1] "A" "A" "A" "A" "A" "A" "A" "A" "A"
> mat
[1] "a" "a" "a" "a" "a" "a" "a" "a" "a"
> mode(pat)
[1] "character"
```

The function “rep” produces a vector with n repeats of its first argument. Since the first argument is a character string, the result is a vector of character strings. The content of a regular vector can either be numerical, logical (“TRUE” or “FALSE”), character strings, raw bytes, or complex numbers.

```
> from.mat <- rbinom(n,1,0.5)
> offspring <- pat
> from.mat==1
[1] FALSE FALSE TRUE TRUE FALSE TRUE FALSE FALSE TRUE
> offspring[from.mat==1]
[1] "A" "A" "A" "A"
> mat[from.mat==1]
[1] "a" "a" "a" "a"
> offspring[from.mat==1] <- mat[from.mat==1]
> offspring
[1] "A" "A" "a" "a" "A" "a" "A" "A" "a"
```

The vector “`from.mat`” is a numerical vector of zeros and ones while the vector “`from.mat==1`” is a logical vector. The index of a vector is given within square brackets. A logical vector can be used in order to select a part of a vector – the part associated with a “TRUE” values for the indexing vector. Compare, for example the indexing vector produced by the third expression above with the indexed vectors given in the subsequent expressions. Finally, note that the assignment in the sixth expression produces a vector of character strings for the segregated alleles according to the randomization produced in the first expression. An alternative approach for selecting a part of a vector is by using a vector of integers for indexing. Using a minus sign will produce the complementary vector. Examine the expressions below. (Recall that the binary operation “`a:b`” produces the vector of integers between `a` and `b`. See the the description of the function “`seq`”.)

```
> 2:6
[1] 2 3 4 5 6
> offspring[2:6]
[1] "A" "a" "a" "A" "a"
> offspring[-(2:6)]
[1] "A" "A" "A" "a"
```

Return to the simulation. We will need to repeat the process of segregation of alleles from a parent to its offspring several times. It is convenient to have a function that conducts this task, instead of writing the same lines of code over and over again. Below we create the appropriate function and store it in an object called “`meiosis`”. Observe the format of a function. It starts with the reserved word “`function`”, followed by a list of its arguments enclosed in parentheses. Next comes the expression that the function applies to the arguments. Instead, one can put between the curly brackets a sequence of expressions to manipulate the arguments. The output of the function is the value of the expression, or it may also be set by the function “`return`”. In our example, if the arguments of the function are the two alleles of the given parent, then the output is the random allele segregated to the offspring. The function “`length`” determines the length of a vector.

```
> meiosis <- function(GF,GM)
+ {
+   from.GM <- rbinom(length(GF),1,0.5)
+   GS <- GF
+   GS[from.GM==1] <- GM[from.GM==1]
+   return(GS)
+ }
> meiosis(pat,mat)
[1] "a" "a" "a" "A" "a" "a" "a" "a" "a" "A"
```

Another special type of vector is called “`list`”. Regular vectors must have all their components of the same type (numerical, logical, character, bites,

or complex). A list, on the other hand, may have any type of object as its component.

```
> n <- 10^5
> model <- list(mu=5,alpha=1,delta=-1,sigma=0.5,allele="A")
> pheno1 <- rnorm(n,model$mu,model$sigma)
> pheno2 <- rnorm(n,model$mu + 2*model$alpha,model$sigma)
> a <- rep("a",n)
> A <- rep("A",n)
> IB1 <- list(pat=a,mat=a,pheno=pheno1)
> IB2 <- list(pat=A,mat=A,pheno=pheno2)
```

The list “model” contains the parameters of the genetic model. Names are assigned to the components of the vector. One alternative for referring to a component of a vector is by its name: “vector.name\$component.name”. (Or, one may use the format: “vector.name[“component.name”]”.) The vectors “pheno1” and “pheno2” store the generated phenotypes of the two inbred lines. Finally, we store the genotype and phenotype information of the two inbred lines as lists, titled “IB1” and “IB2”, respectively.

The function “cross” applies the function “meiosis” in order to simulate a cross between two mice. The first two input arguments are lists, “fa” and “mo”, which contain the genetic information of the two parents. The third argument is a list with the details of the genetic model. (The format “argument=argument.value” may be used in order to assign a default value to the argument. The default value is used unless another value is specifically assigned.) The output of the function is a list with the genetic and phenotypic information of the offspring. Note that the offspring’s “pat” genotype is an allele from the father’s genotype and the offspring’s “mat” genotype is an allele from the mother’s genotype. The object *x* is a vector of integers (0, 1, or 2), *m* is the vector of the offspring’s mean phenotype, and *y* is the vector of expressed phenotypes.

```
> cross <- function(fa,mo,model)
+ {
+   pat <- meiosis(fa$pat,fa$mat)
+   mat <- meiosis(mo$pat,mo$mat)
+   x <- (pat==model$allele)+(mat==model$allele)
+   m <- model$mu + x*model$alpha + (x==1)*model$delta
+   y <- m + model$sigma*rnorm(length(x))
+   return(list(pat=pat,mat=mat,pheno=y))
+ }
```

Note that the function “cross” may use the object “meiosis” even though this object is not implicitly passed as one of the arguments. In general, any existing object may be used inside a function. This is a useful property, but may cause unexpected side effects if applied carelessly.

We apply the function “`cross`” in order to create the F_1 , intercross, and the two types of backcross:

```
> F1 <- cross(IB1,IB2,model)
> BC1 <- cross(IB1,F1,model)
> BC2 <- cross(IB2,F1,model)
> F2 <- cross(F1,F1,model)
```

Since we would like to make several plots of the same format, we create the function “`plot.cross`” in order to facilitate plotting. The input arguments are the name of the cross, which appears in the title of the plot, and a cross list. The output of the function is null. The side effect is the appropriate plot. The expression that produces the plot is very similar to those that were used in the previous plots.

```
> plot.cross <- function(cross.name,cross)
+ {
+   plot(density(cross$pheno),main=paste(cross.name,
+   ":", mean = ",round(mean(cross$pheno),2),
+   ", sd = ",round(sd(cross$pheno),2),sep=""))
+ }
```

The following expressions produce the plots in Fig. 2.2.

```
> op <- par(mfrow=c(2,3))
> plot.cross("IB1",IB1)
> plot.cross("IB2",IB2)
> plot.cross("F1",F1)
> plot.cross("BC1",BC1)
> plot.cross("BC2",BC2)
> plot.cross("F2",F2)
> par(op)
```

The function “`par`” is used in order to set the parameters for high level plotting. We save the current setting of plotting in the object “`op`” and set the plotting region to contain six plots in two rows and three columns. After plotting, the default setting is restored.

Examine the six plots in Fig. 2.2. The first two plots describe the distribution of the phenotype in the two pure inbred strains. Note that the distribution follows the bell shape of the normal distribution. The two distributions look the same, but they have different means. The distribution of the phenotype among the F_1 mice is identical to the distribution among the inbred strain with the low expression, since the genetic effect is recessive. The picture is the same for the $BC1 = aa \times Aa$ backcross. All animals in this cross have at most a single A allele. Genetic variability emerges in the last two plots. Observe that the standard deviation of the phenotype increases from 0.5 to 1 in the case of the intercross and to 1.12 in the case of the backcross. Note that the distribution is no longer normal, but rather a mixture of normals.

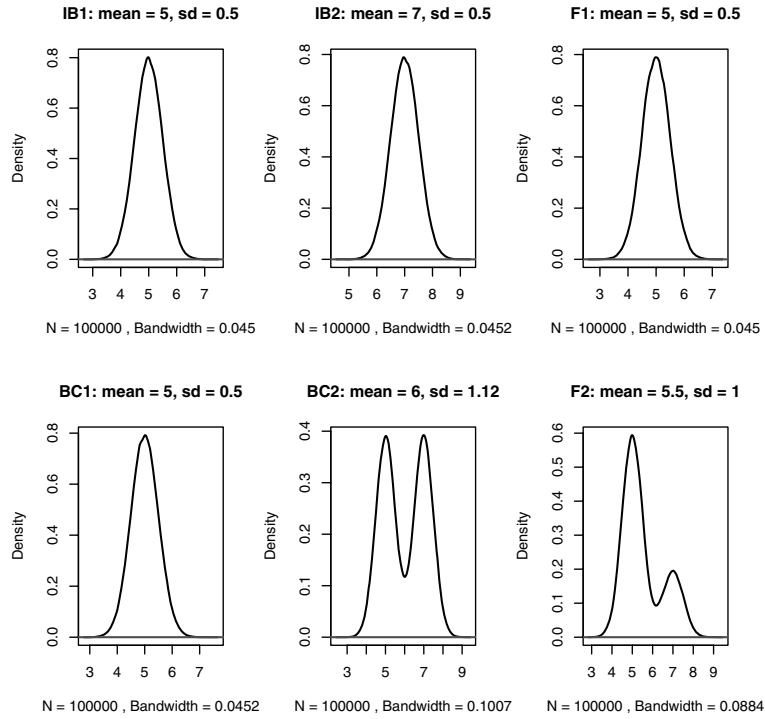


Fig. 2.2. Segregation of the phenotype in crosses

The mixture frequencies are $(1/2, 1/2)$ for the backcross and are $(3/4, 1/4)$ for the intercross.

The investigation of the backcross and the intercross will proceed throughout most of the part of the book that deals with experimental genetics. Other crossing designs will be considered occasionally. In particular, we will deal with *recombinant inbred strains*, which are important resources in genetic mapping. These inbred strains are created by the formation in parallel of several F_2 mating pairs, followed by several generations of inbreeding within each pair. The result is a set of inbred strains, all originating from the initial cross of the two inbred strains. In Chap. 3 we will explain in more detail the genetic population dynamics of the formation of inbred strains and their genetic properties.

2.3 Molecular Genetic Markers

The emergence of modern experimental genetics goes hand in hand with the discovery and the development of the technology for genotyping molecular genetic markers. These markers are specific variations in the sequence of the

DNA molecule. The ability to measure these variations enables the researcher to trace the segregation of segments of genetic material from one generation to the next. The information gathered this way, together with the phenotypic data, is the input used for the statistical analysis. We consider here the two most important types of variations in the context of gene mapping: the *Single Nucleotide Polymorphism* (SNP) markers and the *Simple Sequence Repeat* (SSR) markers.

2.3.1 The SNP Markers

The SNP, as the name suggests, involves variation in the nucleotide at a specific locus. While some in a population of a given species may have, for example, the base *A* at a given locus on a particular strand of DNA, others may have *T* at the very same location in the DNA sequence. Although, in principle, SNPs may have four distinct alleles, in practice they are typically bi-allelic. Another, less frequent type of a SNP, is a deletion, i.e., the absence of a given base pair.

The SNPs are the most abundant form of variation found in the genomes of mammals. More than a million such variations have been mapped in the last few years in the human genome. A similar effort, currently under way for the mouse genome and for other genomes, is sure to produce similar numbers for use in experimental genetics in the near future.

Current technologies for genotyping involve multiplying the region of the variation using the *Polymerase Chain Reaction* (PCR). Via this simple but miraculous laboratory process, a tube that originally contained a relatively small number of very long and complex genomic DNA molecules, ends up containing a huge number of copies of a small segment of the molecule about the point of variation. Thus, instead of having to determine the color of a needle in a stack of hay, the problem becomes that of determining the color of the needle in a stack of identical needles.

Modern technologies involve various methods of partial sequencing of the polymorphic locus, different approaches of tagging the locus, and signaling to an electro-optic or other type of a detection device. New technologies emerge almost daily, pushing down the price and increasing the rate of the genotyping of SNPs.

2.3.2 The SSR Markers

Undoubtedly, SNPs will have a significant role in the future of experimental genetics, including the mouse model. However, until that day comes, SSR markers are still the most important form of variation for mapping in experimental genetics (and in family-based linkage analysis in humans).

An SSR – more commonly known under the nickname “microsatellite” – is a genomic element that consists of a mono-, bi-, tri-, or tetrameric sequence repeated (hence the name) a number of times that varies from one individual

to another. These elements are very polymorphic. They also tend to mutate relatively rapidly. However, this is not a major concern in experimental genetics and in linkage analysis, which typically involve tracing the segregation process only a limited number of generations.

The alleles of an SSR marker are determined by the length of the SSR element. Thus, applying PCR to segments containing the SSR element will yield products that vary in length according to number of tandem repeats in the alleles. Separating these products by length, either by the use of gel electrophoresis, or by the use of more sophisticated capillary-based systems, gives a direct read of the different alleles. Tens of thousands of such SSR markers have been identified and mapped. Commercial kits for their genotyping make this tool handy for even the most modest genetic laboratories.

2.4 Bibliographical Comments

The regression model of this chapter originated in the pioneering paper of Fisher [30]. A systematic development of a general version of this model is given by Kempthorne [44]. Clearly written expositions appear in Falconer and Mackay [27], Crow and Kimura [16], Lander and Botstein [47], and Lynch and Walsh [51], among others.

Problems

For the following problems, assume that e has a standard normal distribution.

2.1. Simulate the distribution of y for a backcross design. Do the same for an intercross design. Consider various levels of α and δ .

2.2. Consider the two major genes additive model:

$$y = \mu + \alpha_1 x_1 + \alpha_2 x_2 + e, \quad (2.8)$$

where x_i denote the number of A_i alleles at locus i ($i = 1, 2$). Assume the genes corresponding to x_1 and x_2 lie on two different chromosomes, so by Mendel's laws x_1 and x_2 are independent.

(a) Investigate the distribution of the phenotype for various values of the model parameters (including the probabilities p_i of the allele A_i).

(b) Assume that the indicated genes are the only genes contributing to the trait, so e can be regarded as the environmental effect E . Find expressions for σ_G^2 and σ_y^2 . (It will be helpful to rewrite the model as in (2.3).)

(c) Assume in addition that the parental strains are inbred. How would you estimate H^2 if you know the phenotypes of samples from both parental strains and from the intercross?

(d) Extend the model to include k independent genes, $k \geq 2$. Assume that $\alpha_i = \alpha$, for all i and that the parental strains are inbred. Can you figure out a way to estimate k and α from phenotypic data involving both parental strains and the intercross progeny?

2.3. Show that (2.3) follows from (2.2). Verify that the two terms involving x_M and x_F on the right-hand side of (2.3) are uncorrelated. Hence verify (2.4) and (2.5). Hint: To facilitate algebraic manipulations, it may be helpful to observe that $|x_M - x_F| = x_M + x_F - 2x_Mx_F$.

2.4. The model (2.3) is customarily written in the somewhat different form

$$y = m + \tilde{\alpha}(x - 2p) + \tilde{\delta}[I_{\{x=1\}} - 2p(1-p) - (1-2p)(x-2p)] + e,$$

where $\tilde{\alpha} = \alpha + (1-2p)\delta$ and $\tilde{\delta} = -2\delta$. Show that this is the same as (2.3). The form given in (2.3) seems slightly easier to manipulate computationally and illustrates that what geneticists call “dominance” is exactly what statisticians call “interaction,” in this case the interaction of the allele inherited from the mother with that inherited from the father.

2.5. Generalize the model of Prob. 2.2 to include interaction between loci, as follows: Starting from $y = \mu + \alpha_1x_1 + \alpha_2x_2 + \gamma x_1x_2 + e$, where as above the two loci are assumed to lie on different chromosomes, re-write the model in the form:

$$y = m + \tilde{\alpha}_1(x_1 - 2p_1) + \tilde{\alpha}_2(x_2 - 2p_2) + \gamma(x_1 - 2p_1)(x_2 - 2p_2) + e.$$

What are $\tilde{\alpha}_1$ and $\tilde{\alpha}_2$? Find an expression for σ_y^2 . The term $\gamma^2 p_1(1-p_1)p_2(1-p_2)$ is called the interaction variance, or more precisely the additive-additive interaction variance to distinguish this form of interaction from other possibilities.

The Statistics of Gene Mapping

Siegmund, D.; Yakir, B.

2007, XX, 334 p., Hardcover

ISBN: 978-0-387-49684-9