Analysis of quantitative trait loci

7

Many traits are **polygenic**, that is, determined by the cumulative contributions of a large number of genes. Furthermore, many of these traits are capable of being measured on a scale at least as strong as the interval scale; such traits are **quantitative traits**. If a polygenic trait Y is quantitative and the contributions from its governing genes are assumed to combine additively, we have the **additive genetic model**, presented earlier as eqn (8.1):

$$Y = \sum_{\lambda \in \mathbb{G}} \left(\Upsilon_\lambda^{\female} + \Upsilon_\lambda^{\male} \right) + \Upsilon_E \, . \tag{7.1}$$

Here Y is the value of a trait of interest, $\mathbb{G}$ is the set of loci that exert an influence on this trait and λ is a locus index ranging over this set. The contributions from the maternal and the paternal chromosome at locus λ are denoted as $\Upsilon_\lambda^{\female}$ and $\Upsilon_\lambda^{\male}$, respectively. The sum (over $\mathbb{G}$) of these contributions makes up the additive genetic component Υ_A to which an environmental influence Υ_E is added.[1] The latter is usually assumed to satisfy $\mathbb{E}[\Upsilon_E] = 0$ and is also often taken to follow the normal distribution, with a nod to the Central Limit Theorem.

It cannot be overemphasised that this is just a model. We need only reflect on the action of genes through the central dogma and an organism's ontology, which involves functional biology at all grades from molecule to ecosystem to conclude that an interaction between genes that is precisely additive would be nothing short of a miracle. Moreover, examples abound of systems where the interaction is markedly non-linear. This leaves a wide class of quantitative traits in which the additivity assumption gives a reasonable approximation. The best attitude to take is that the additive genetic model is a convenient default position, which often proves useful, while we remain mindful of its limited applicability.[2]

Adopting the additively genetic model as our standard framework, the question of determining its parameters arises immediately. We ask how many loci are involved (i.e. can we estimate $|\mathbb{G}|$?) What are the values Υ_λ associated with the alleles at each locus? And where are these **quantitative trait loci** (QTL) within the genome—can be construct a chromosomal map that shows where the loci indexed by λ are located? This estimation problem is important not just for its own sake, as a mathematical exercise, but also because its solution, where possible, represents a first step in linking genes to traits.

[1] This is "environmental" in the widest sense of the word, in effect comprising all non-genetic influences impinging on the organism from the zygote stage onwards.

[2] The assumption of additivity is explored un more depth in Section 10.2.

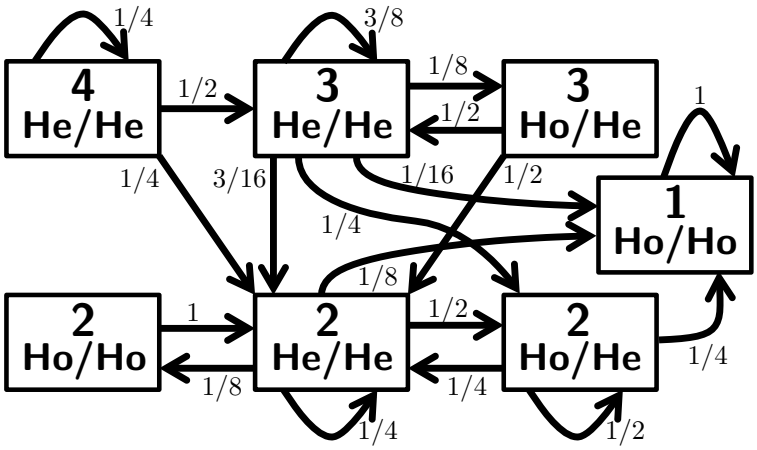


Fig. 7.1 Inbreeding dynamics. Boxes correspond to the genetic status of the breeding pair at a particular locus of interest: for instance, **4 He/He** means that the pair has 4 distinct alleles between them at that locus, and both individuals are heterozygotes with respect to the locus. Transition arrows point to the genetic status of the breeding pair in the next generation, chosen from among the offspring. The arrows are labelled with the probability that this particular transition will occur.

7.1 Recombinant distributions

Although QTL analysis is not difficult in principle, it can be thorny in practice and the complications tend to obscure the principles somewhat. Therefore we will develop the ideas starting from highly idealised cases.

An important resource for controlled genetic experiments is a population of individuals that can be assumed to be homozygous with respect to all the loci of interest. In fact, we will need to dispose of several such lines, which between them represent the alleles of interest. This can be accomplished by **inbreeding**.

7.1.1 Dynamics of inbreeding

The idea of inbreeding is simple: from every set of sibling offspring, choose a brother-sister breeding pair, and repeat this with the offspring they produce. In selective breeding, genetic homogenisation is to be expected, at least for the loci connected to the traits that the breeder is selecting for.[3] But even in the loci not strongly correlated to the traits to which the breeder pays attention, generations of inbreeding will result in thoroughly homozygous individuals.

The reason that inbreeding has this effect is contained in the Markov graph depicted in Fig. 7.1. For autosomal loci, the breeding pair has four copies of the locus between them, and more than one allele may be carried by this foursome in any of the generations. However, only the state with a breeding pair that is homozygous for the same allele is recurrent, and all the other states are transient.[4] Thus, in the long run the probability that the line will have just one allele on all copies of the locus tends to one. This is shown in Fig. 7.2, which shows the probabilities down

[4] Derive the transition probabilities that label the arrows in Fig. 7.1 with the aid of breeding diagrams.

[3] This is the usual way of putting things: organisms are thought of as "having" a large number of traits, some or all of which may be observed or monitored. In Chapter 10, we shall see that this way of thinking engenders several perplexities which are avoided if we think of a trait as a "probe" that is defined by the observation. In human-operated breeding, the issue is somewhat less pressing.

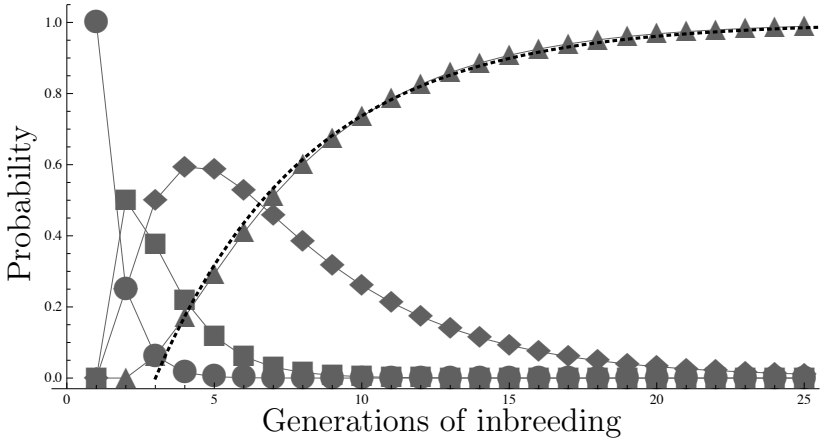


Fig. 7.2 Probabilities of four (circles), three (squares), two (diamonds) alleles, and of a single allele (triangles), remaining in the genetic pool as a function of the number of generations of inbreeding. The dashed line is a graph of the simple approximation
$1 - \exp\{-(n-3)/(3+\sqrt{5})\}$.

the breeding generations starting from the "worse case" which is four distinct alleles spread over the homologous sites on the breeding pair's chromosomes. The 4-allele probability reduces quite quickly, whereas the 2-allele probability lingers for quite a while. Inbreeding therefore has to be pursued for dozens of generations at a minimum, a criterion that will be more time-consuming and costly to satisfy for large organism with a long generation time.[5]

Using the largest eigenvalue of the transition matrix[6] among those that are smaller than 1, we can propose a simple approximation for $p_n(1)$, the probability of having a single allele after n generations of inbreeding:

$$p_n(1) = 1 - \exp\{-(n-3)/(3+\sqrt{5})\} \tag{7.2}$$

a graph of which is shown as a dashed line in Fig. 7.2.

A corollary of this analysis is that, as far as genetic homogeneity is concerned, there is no need for directed selection of the breeding pair guided by trait values. Indeed, since the objective is to have a number of inbred lines that represent the allelic variability at all loci between them, it would be better to select, by random allotment, as many breeding pairs from every generation as the budget will allow.[7] Meeting up of recessive alleles may result in some lines that are sickly, for instance susceptible to infections or spontaneously developing conditions such as (in mammals) hypertension or diabetes. Whereas it may be costly to maintain lines that require special care, and there is an ethical question about the breeding of organisms that are born to be sick, such lines are much sought-after as "models" of similar conditions in humans.

If crossing-over did not occur, the analysis presented here would apply to entire chromosomes. The effect of crossing-over is to mix up the chromosomes, so that the loci behave quasi-indendently.[8] Moreover, **unequal crossing-over** introduces additional variability as a result of errors in alignment. In coding regions, such gene mispairing may result in gene duplications, but more frequently will give rise to lethal mutant alleles (which do not contribute to the variability that the inbreeding

5 There are alleles that have a lethal effect when appearing homozygously, or which may render the organism less likely to be picked for the breeding pair by the breeder. Modify the Markov graph of Fig. 7.1 to account for this possibility.

6 This is a 7×7 matrix. Write it down using the probabilities states in Fig. 7.1. Write down the eigenvector corresponding to the eigenvalue 1. Use computer algebra to obtain the other eigenvalues; which one was used to derive eqn (7.2), and why?

7 This is more crucial for the earlier generations—why?

8 See Exercise 7.1.

is intended to eliminate). In non-coding regions, misregistrations may generate considerable variability; as we shall see in Section 7.2, this kind of variability among the inbred lines is actually a great boon. On the other hand, closely linked loci will still tend to homogenise together.

7.1.2 One locus

We now return to the general question of characterising the contributions made by the loci in $\mathbb{G}$ to a trait of interest Y, having at our disposal a repertoire of inbred lines. We might be fortunate in finding, among these lines, a pair of lines which differ only in one locus. In practice this case is unlikely to arise, but it is a nice point of departure for our analysis.

Suppose that at this unique locus λ two alleles may be found, α_1 and α_2. Observing the trait value Y in various individuals belonging to both lines, we can ascertain whether a difference between these lines is statistically detectable. Equivalently, we can obtain interval estimates for the quantity

$$\Delta\Upsilon_\lambda = \Upsilon(\alpha_2) - \Upsilon(\alpha_1)\ . \tag{7.3}$$

This quantity is related to the average observed value of the trait value in the parental lines by:[9]

$$\mathbb{E}[Y(\alpha_2\alpha_2)] - \mathbb{E}[Y(\alpha_1\alpha_1)] = 2\Delta\Upsilon_\lambda\ . \tag{7.4}$$

A cross of $\alpha_1\alpha_1$ individuals with $\alpha_2\alpha_2$ individuals results in heterozygous offspring F_1 with genotype $\alpha_1\alpha_2$. For the expectation at the phenotypic level is as follows:[10]

$$\mathbb{E}[Y(F_1)] = \mathbb{E}[Y(\alpha_1\alpha_2)] = \frac{\mathbb{E}[Y(P_1)] + \mathbb{E}[Y(P_2)]}{2} \tag{7.5}$$

where P_1 and P_2 denote the parental lines, that is $\mathbb{E}[Y(P_i)] = \mathbb{E}[Y(\alpha_i\alpha_i)]$ for $i = 1, 2$. This is where the additivity assumption faces its first test: if we do not find that eqn (7.5) holds (to within the limits of statistical detectability) this failure may prompt us to reject the assumption of additive effects at locus λ for the trait Y under consideration.

7.1.3 Two loci

Suppose, next, that we have two parent lines[11] P_1 and P_2 that differ in precisely two of the loci in $\mathbb{G}$. At the first locus λ there are two alleles, α_1 and α_2, and at the second locus λ' there are two alleles, β_1 and β_2. Lest the analysis immediately reduces to the previous case, we will take it as given that both loci have been found to exert a statistically detectable effect on the trait value Y of interest.

The genotypes of the parents are $\alpha_1\alpha_1\beta_1\beta_1$ and $\alpha_2\alpha_2\beta_2\beta_2$.[12] Parent i produces gametes of haplotype $\alpha_i\beta_i$; crossing-over events may occur but will not affect the outcome since the parents are homozygotes. As a

[9] Verify this.

[10] Show that eqn (7.5) follows from the additive genetic model.

[11] QTL jargon can be confusing: these inbred homozygous lines, from which experiments and analysis depart, are referred to throughout as the "parents" even though we will be discussing their grandchildren, great-grandchildren, and so forth.

[12] There is no loss of generality here: we simply *name* the alleles for the parents in which they are present. We have to be careful, though, when we repeat the experiment with another inbred line which may have genotype $\alpha_1\alpha_1\beta_2\beta_2$ in terms of the present notation.

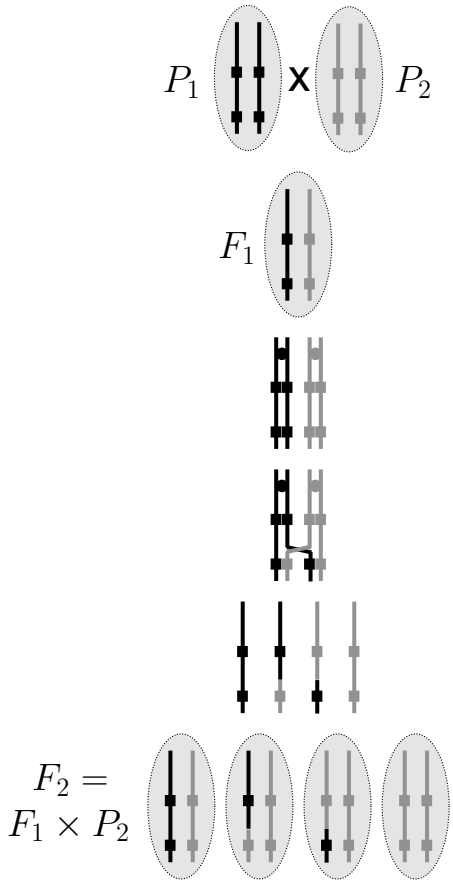


Fig. 7.3 Formation of recombinants through back-crosses of nearly homozygous parental lines. The top row depicts schematically the nuclei of the two parent lines, which differ in two loci, and are both homozygous. Crossing these lines, one obtains offspring that is heterozygous for these two loci (and is also homozygous for any other locus affecting the trait under study, by assumption). The offspring produces two kinds of gametes, as a result of crossing-over during meiosis I. Thus, crossing back the offspring with one of the parent lines, four kinds of offspring are obtained, one of which resembles the F_1 offspring, one of which resembles the parent line, and two of which are recombinants which are intermediate in trait value by the additive genetic model.

result, the offspring (F_1) produced from a $P_1 \times P_2$ cross will have genotype $\alpha_1\alpha_2\beta_1\beta_2$ and these organisms produce gametes with four genotypes: $\alpha_1\beta_1$, $\alpha_1\beta_2$, $\alpha_2\beta_1$, and $\alpha_2\beta_2$. Of these, the middle two are of special interest, since they are **recombinants**. When λ and λ' are located on separate chromosomes and their assortment during meiosis is independent, the frequencies of all four types of gametes will be equal (i.e. $1/4$).[13] On the other hand, when they are on the same chromosome and no cross-over occurs between them, there will be no recombinants at all and the two parental genotypes occur each at frequency $1/2$. Some degree of cross-over will occur, however, so in general we should reckon with a recombination fraction as a real[14] number in the interval $[0, 1]$ (or rather the interval $[0, \frac{1}{2}]$; see Exercise 7.9).

The extent of cross-over will be a monotone increasing function of the physical distance on the chromosome between the loci. However, this relationship should not be thought of as a simple proportionality because some parts of the chromosome are more prone to a cross-over event than others: whereas some regions may be recombination "hot spots", others have a suppressed rate of cross-over. The centromer is an example of

[13] We assume that none of the alleles involved affects segregation or success in meiosis.

[14] A real number, or a rational number? Why would do we prefer to work with real numbers?

such a "cold spot."[15]

Crossing the F_1 individuals with an organism from parent line P_1 (this is known as a **back-cross**), we obtain an F_2 ("grandchildren") comprising four types of individuals: one with the P_1 genotype, one with the F_1 genotype and two with the recombinant genotypes $\alpha_1\alpha_2\beta_1\beta_1$ and $\alpha_1\alpha_1\beta_1\beta_2$. The entire procedure is depicted schematically in Fig. 7.3. We could equally well have back-crossed with parent P_2.

As before, we assume that in all these organisms we are able to observe the trait value Y. Although we are not in a position to obtain Υ-values,[16] we can estimate $\Delta\Upsilon_\lambda$ and $\Delta\Upsilon_{\lambda'}$. We have the following identities:[17]

$$\begin{aligned}\Delta\Upsilon_\lambda = \Upsilon(\alpha_2) - \Upsilon(\alpha_1) &= \mathbb{E}[Y(\alpha_1\alpha_2\beta_1\beta_1)] - \mathbb{E}[Y(\alpha_1\alpha_1\beta_1\beta_1)] \\ &= \mathbb{E}[Y(\alpha_1\alpha_2\beta_1\beta_1)] - \mathbb{E}[Y(P_1)] \qquad (7.6)\\ \Delta\Upsilon_{\lambda'} = \Upsilon(\beta_2) - \Upsilon(\beta_1) &= \mathbb{E}[Y(\alpha_1\alpha_1\beta_1\beta_2)] - \mathbb{E}[Y(\alpha_1\alpha_1\beta_1\beta_1)] \\ &= \mathbb{E}[Y(\alpha_1\alpha_1\beta_1\beta_2)] - \mathbb{E}[Y(P_1)] \qquad (7.7)\end{aligned}$$

where the expectations can be estimated by averaging the observed trait values in the respective groups of genotypes. It may happen that $\Delta\Upsilon_\lambda = -\Delta\Upsilon_{\lambda'}$. In that case the additive genetic model gives $\mathbb{E}[Y(P_1)] = \mathbb{E}[Y(P_2)] = \mathbb{E}[Y(F_1)]$.[18] The genetic difference between the parent lines only reveals itself phenotypically in the F_2 generation, since one of the recombinant trait values will be larger (on average) than this common value, and one will be lower.

Additivity and epistasis

To assess the validity of the additivity assumption, we can use the following identity:

$$\Delta\Upsilon_\lambda + \Delta\Upsilon_{\lambda'} = \mathbb{E}[Y(F_1)] - \mathbb{E}[Y(P_1)]\ . \qquad (7.8)$$

The quantities $\Delta\Upsilon_\lambda$ and $\Delta\Upsilon_{\lambda'}$ as defined by eqns (7.6) and (7.7) are defined regardless of the validity of the additivity assumption, and may be calculated from the data. The same is true of the right-hand side of eqn (7.8).[19] Hence we have an additional check.

At first glance, it might seem superfluous to carry out a back-cross with the other parent, P_2. After all, we should expect to glean the same information. However, if the additivity assumption is not satisfied we can gauge the extent of the deviation by comparing the results obtained when back-crossing with different parents. Let $\Delta\Upsilon_\lambda^{[p]}$ denote the effect estimated with parent p. Then the **dominance deviation** is defined as follows:[20]

$$\varepsilon_\lambda = \frac{\Delta\Upsilon_\lambda^{[1]} - \Delta\Upsilon_\lambda^{[2]}}{2}\ . \qquad (7.9)$$

This quantity ε_λ can be regarded as a correction term if the underlying additive effect is taken to be the average of the observed effects:[21]

$$\Delta\Upsilon_\lambda = \frac{\Delta\Upsilon_\lambda^{[1]} + \Delta\Upsilon_\lambda^{[2]}}{2}\ . \qquad (7.10)$$

[15] See Exercise 7.4.

[16] We can do the next best thing, though; see Chapter 10.

[17] What assumptions do we need to make about the statistical distribution of Υ_E? What can be done in practice to ensure that these assumptions are likely to be satisfied?

[18] Why?

[19] Verify eqn (7.8) by writing out F_1 and P_1 as genotypes.

[20] Verify that $\varepsilon_\lambda = 0$ when the additive genetic model applies.

[21] Check that

$$\Delta\Upsilon_\lambda^{[1]} = \Delta\Upsilon_\lambda + \varepsilon_\lambda\ .$$

Give a similar formula for $\Delta\Upsilon_\lambda^{[2]}$.

In what follows, we shall always assume that the protocol is followed both ways and the estimated $\Delta\Upsilon$s averaged over the parental lines.

The dominance deviation may be regarded as measure of the non-linear interaction between the alleles situated at the same locus on the two homologous chromosomes. There may also be non-linear interactions between the alleles at two or more different loci. The term **epistasis** refers to a interaction between alleles, such that the effect of one locus on the trait (or on the organism's fitness) depends on the alleles at one or more other loci. Stated in such vague terms, it is just a truism, because the interdependency of genes[22] no allele can exert its effect independently of the genomic context in which it finds itself.

[22] By way of the Central Dogma, development (ontogeny), and physiology (functional biology).

If we take the "effect" of locus to mean its additive contribution, epistasis would refer to any inter-locus iteration beyond the scope of the additive model. This suggests that we could set up a hierarchy of models: at the basis is the additive model, on top of which we add corrections due to interactions between pairs of loci (second-order, or "two-point" interactions), to which we add another layer of corrections due to interactions between trios of loci (third-order, or "three-point" interactions), and so on, at least in principle. This idea is developed further in Section 10.2.1.

However, an attempt to define epistasis in terms of non-additive effects suffers from a conceptual flaw since the particular way in which a trait of interest has been quantified affects the applicability of the additive genetic model. Consider for instance a trait Y which is described perfectly by the additive model. Now define another trait variable Y' as $\ln Y$. Even though Y and Y' are, strictly speaking, distinct traits, it would be more natural to regard them as different ways of representing the same underlying biological reality. This prompts us to say that the additive model applies, in a general sense, if it adequately[23] describe a trait variable of interest or any monotone transformation thereof.

[23] As always, "adequately" depends on the quality of approximation that is sufficient for the purpose at hand.

"Hard" epistasis occurs when even the additive model together with a monotone transformation does not suffice. This occurs when the manifestation of a trait depends in an all-or-none fashion on a certain combination over two or more loci being present.[24] One example is that of **mimicri**, in which an organism's morphology confounds predators by a sufficiently convincing resemblance (as seen through the eyes of the predator) to a species that the predator has learned to avoid as toxic or distasteful.[25] The resemblance requires a "concordance" between several phenotypic features, which amounts again to an "exclusive or"-type interaction between alleles at several loci.

[24] Some authors define epistasis more specifically as the situation where expression of one gene obscures the phenotypic effects of another gene.

Another classical example is **heterostyly** in plants, where the male

[25] In **Batesian** mimicry, the "pretender" actually lacks the properties that discourage the predator and is effectively banking on the predator's experience with the noxious model. In **Müllerian mimicry**, two or more noxious species resemble one another; the selective advantage then resides in the fact that the burden of instructing predators, which has to be repeated for each individual member of the predating species, is spread among more individuals that partake in the pool of mutual resemblance, thus reducing the average risk to individual mimics. This explains the existence of universal warning colours, such as the yellow/black, red/black, and red/yellow/black liveries so characteristic of poisonous or stinging snakes, amphibians, hymenoptera and so on.

Fig. 7.4 Heterostyly as an example of epistasis. Shown are four partially dissected *Primula* flowers. In the top row, the filament of the stamen (male reproductive part) and the style of pistil (female reproductive part) have differing lengths (heterostyly), giving rise to a self-sterile (cross-pollinating) phenotype, whereas in the bottom row, filament and style have the same length (homostyly), giving rise to a self-fertilising phenotype which has lower fitness but which can arise from heterostylous parents through recombination. The heterostylous phenotypes are known as **pin** when the style is longer and **thrum** when the filaments are longer. The heterostylous and homostylous phenotypes are related to the loci governing the lengths of the filament and the style according to an *exclusive or (XOR)* relationship, which is an example of a "hard" non-linearity.

and female parts have differing lengths (Fig. 7.4). This difference precludes self-fertilisation and promotes cross-pollination. However, crossovers can result in **homostylous** flowers which have male and female parts of equal length, a type that can self-fertilise. In this case too, the phenotypes arise through an "exclusive or"-type interaction between alleles at different loci. Such interactions are not readily reconciled with the additivity assumption.[26] Incidentally, the lengths of the male and female parts are quantitative traits that have been successfully analysed using the QTL approach.

[26]The problem is studied in more detail in Section 10.2.3.

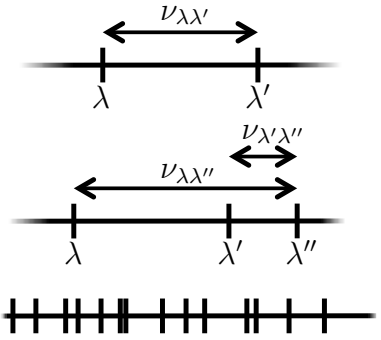


Fig. 7.5 The recombination fractions between loci are used to lay out a tentative map.

Estimating the parameters

The fraction of recombinants ν is an indication of the physical relationship of the loci: when $\nu < \frac{1}{2}$, we putatively put the two loci on a chromosomal map. We draw the chromosomal map as a straight line and see to it that the difference between the loci on the map is proportional to ν. Initially, we are quite unsure just where the loci are on the chromosome—towards one of the telomeres, or closer to the centre, unless the recombination fraction is quite close to $\frac{1}{2}$. However, as we find out recombination fractions for more loci, we start filling out a line segment corresponding to the entire chromosome (Fig. 7.5).

In doing so, we ignore the fact that physical distance is not directly

proportional to recombination fraction. An actual physical map that places the loci at the correct base-distances from one another will stand in a a monotone bijective relationship with the map we draw. The maximal distance on our ν-based map is $\frac{1}{2}$; our map distance is dimensionless, or one might say expressed in a "probability unit" which traditionally is called the **morgan**; thus the longest distance on a single chromosome is 0.5 morgan. Distances between loci are expressed in centimorgan (cM).

Unless we bring molecular-genetic techniques to bear, we have no way of knowing which of the recombinants is which. Such ignorance is not a dramatic drawback since λ, λ', and so on have thus far merely served as interchangeable labels for the loci. Moreover, as we repeat the crossing and back-crossing experiment with more and more lines, we would encounter repeated values for the differences $\Delta\Upsilon$ which allows us to postulate common loci between the parental lines and start building up a chromosomal map.

We would be able to partition the F_2 offspring into four groups and apply eqns (7.6)—(7.8) directly, if we were able to identify genotypes. To be sure, there are techniques to accomplish just this, as we shall see below in Section 7.2; but let us first consider how much inroads we can make when such means are not at out disposal, if only to appreciate why the techniques we consider in Section 7.2 represent such a formidable leap forward.

We can avoid the need to assign genotypes if we adopt a maximum-likelihood approach. The likelihood of the observations can be expressed as follows:

$$\begin{aligned}\mathbb{L}[\mu_{P_1}, \sigma^2, \Delta\Upsilon_\lambda, \Delta\Upsilon_{\lambda'}, \nu; \{Y_1, \ldots, Y_N\}] = \\ \prod_{i=1}^{N} \left\{ \frac{1-\nu}{2} \left(f(Y_i; \mu_{P_1}, \sigma^2) + f(Y_i; \mu_{P_1} + \Delta\Upsilon_\lambda + \Delta\Upsilon_{\lambda'}, \sigma^2) \right) + \right. \\ \left. \frac{\nu}{2} \left(f(Y_i; \mu_{P_1} + \Delta\Upsilon_\lambda, \sigma^2) + f(Y_i; \mu_{P_1} + \Delta\Upsilon_{\lambda'}, \sigma^2) \right) \right\} \quad (7.11)\end{aligned}$$

where N is the number of observed F_2 offspring, with trait value Y_i for the ith individual, $f(\cdot; \mu, \sigma)$ is an appropriately chosen probability density function (pdf) with mean μ and variance σ^2. The parameters of most interest are the recombination fraction ν and the loci-associated differences $\Delta\Upsilon_\lambda$ and $\Delta\Upsilon_{\lambda'}$.[27]

Of the F_2, a proportion $(1-\nu)$ will have the phenotype of a parental line. Of these, one half have the phenotype of P_1, centering the distribution of Y among these individuals at μ_{P_1}, whereas the expected trait value for the other half is the P_2-mean which can be written as $\mu_{P_1} + \Delta\Upsilon_\lambda + \Delta\Upsilon_{\lambda'}$. Furthermore, a proportion ν of the F_2 will have a recombinant phenotype, half of them with a mean that deviates from the P_1 mean at locus λ, giving a mean $\mu_{P_1} + \Delta\Upsilon_\lambda$, and half of them deviating from P_1 at λ', giving $\mu_{P_1} + \Delta\Upsilon_{\lambda'}$. Combining all these terms, we arrive at the above expression for the likelihood.[28]

The MLE estimates for these parameters are found by maximising the likelihood $\mathbb{L}$. In practice, we maximise the log-likelihood ($\ln \mathbb{L}$) by

[27] Observe that

$$\mu_{P_2} = \mu_{P_1} + \Delta\Upsilon_\lambda + \Delta\Upsilon_{\lambda'} .$$

[28] Observe that the expression between curly braces in eqn (7.11) is itself a pdf of a mixed distribution (see Section 4.2.2 for a definition of the latter).

numerical means.[29] Whatever algorithm we use for this, it will require initial estimates. A suitable initial estimates for μ_{P_1} is $\langle Y(P_1)\rangle$, the averaged observed trait value in parent line P_1. Similarly, an initial estimate for both $\Delta\Upsilon_\lambda$ and $\Delta\Upsilon_{\lambda'}$ is given by $(\langle Y(F_1)\rangle - \langle Y(P_1)\rangle)/2$. For σ the pooled sample standard deviation in the P_1 and F_1 populations furnishes an initial estimate. Finally, $\frac{1}{2}$ should be a good initial guess for ν.

If we exchange the MLE values for $\Delta\Upsilon_\lambda$ and $\Delta\Upsilon_{\lambda'}$, we obtain an equally good likelihood—the loci are indistinguishable by the present method. This is not a major stumbling block, since the salient properties of the loci are their contributions (or the Δ-values, to be more precise) and their "distance" in terms of ν. Care should be taken, however, when we compare the results of repetitions of the entire experiment with different pairs of lines. A concordance of ν-values, as illustrated schematically in Fig. 7.5, can then be used to work out which loci must be identified between such experiments.

7.1.4 More than two loci

The case of three or more loci does not present any new conceptual difficulties. Consider first the case of three loci and suppose that the staring material consists of two lines, P_1 and P_1, that differ in precisely *three* of the loci in $\mathbb{G}$. Let the loci be λ, λ', and λ''. Offspring F_1 and F_2 are created as before (back-crossing F_1 with the P_1 parent to obtain F_2); the latter now comprises $2^3 = 8$ genotypes, of which two are P_1 and F_1 and six are recombinants. The likelihood is given by the following expression:

$$\begin{aligned}\mathbb{L}[\mu_{P_1}, \sigma^2, \Delta\Upsilon_\lambda, \Delta\Upsilon_{\lambda'}, \Delta\Upsilon_{\lambda''}, \nu_{ab}, \nu_{bc}, \nu_{ac}; \{Y_1, \ldots, Y_N\}] = \\ \frac{1}{2}\prod_{i=1}^{N}\big\{(1-\nu_{ab}-\nu_{bc}-\nu_{ac})\big(f(Y_i;\mu_{P_1},\sigma^2)+ \\ f(Y_i;\mu_{P_1}+\Delta\Upsilon_\lambda+\Delta\Upsilon_{\lambda'}+\Delta\Upsilon_{\lambda''},\sigma^2)\big)+ \\ \nu_{ab}\big(f(Y_i;\mu_{P_1}+\Delta\Upsilon_\lambda,\sigma^2)+f(Y_i;\mu_{P_1}+\Delta\Upsilon_{\lambda'}+\Delta\Upsilon_{\lambda''},\sigma^2)\big)+ \\ \nu_{bc}\big(f(Y_i;\mu_{P_1}+\Delta\Upsilon_\lambda+\Delta\Upsilon_{\lambda'},\sigma^2)+f(Y_i;\mu_{P_1}+\Delta\Upsilon_{\lambda''},\sigma^2)\big)+ \\ \nu_{ac}\big(f(Y_i;\mu_{P_1}+\Delta\Upsilon_\lambda+\Delta\Upsilon_{\lambda''},\sigma^2)+f(Y_i;\mu_{P_1}+\Delta\Upsilon_{\lambda'},\sigma^2)\big)\big\} \quad (7.12)\end{aligned}$$

which is entirely analogous to eqn (7.12) except that there are a few more parameters and terms.[30] The pdf of the mixed distribution is motivated diagrammatically in Fig. 7.6.

The total number of observations N and the width of the peaks σ determine how readily the parameter estimates can be extracted from the data. The larger N and the smaller σ, the less likely the estimation algorithm is expected to throw up problems.[31] For more than three loci, writing down the mixed pdf follows the same principles, but the demands on N and σ as regards parameter identifiability rapidly become a great deal more stringent. If there are n loci in play, there will be 2^n peaks

[29] The factor $\frac{1}{2}$ in eqn (7.11) was included for the emphasise the connection to the counting argument used to derive the formula; it can be dropped from the calculations.

[30] If the three loci lie on three distinct chromosomes, we would expect $\nu_{ac} = \nu_{ab} = \nu_{bc} = \frac{1}{4}$. If λ' and λ'' lie on the same chromosome which is distinct from the chromosome with λ, we should find $\nu_{ab}(1-\nu)/2$ and $\nu_{bc} = \nu_{ac} = \nu/2$, where ν is a common parameter. Explain these expectations and write down similar formulas for the cases where λ' or λ'' is the unlinked chromosome. Explain how the generalised likelihood ratio principle (Section 5.2.2) can be used to test these hypotheses.

[31] Under no circumstance should N be smaller than the total number of parameters.

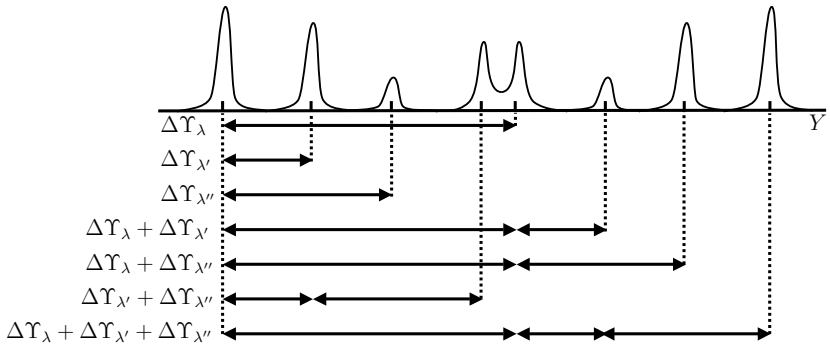


Fig. 7.6 Probability density function of a trait Y in the F_2 (back-cross) population with peaks at the parental phenotypes as well as six recombinants. The contributions from loci λ, λ', and λ'' are indicated. Recombinant peaks occur in pairs of equal probability mass.

which will inevitably start merging and turn the distribution in a quasi-uniform "smear" in which many or most of the individual peaks no longer stand out.[32]

[32] The smaller σ, the larger n can be before this smear-effect kicks in. But the more relevant point is that in typical applications, this will happen at moderate n far below the number of contributing loci.

7.2 Genetic markers

One of the main difficulties with the approach outlined in the preceding section is that in general it is not be known in advance at how many loci the parent lines differ with respect to the loci in the set $\mathbb{G}$ (the set of loci governing the trait of interest). We could just fit the mixed distribution, first on the assumption that $n = 1$, then assuming $n = 2$, $n = 3$, and so on. At each stage, we can apply the generalised likelihood principle[33] to ascertain whether the increment of n by 1 adds a statistically detectable improvement. In most cases, parameter identifiability will become limiting before there is firm evidence that n is large enough to account for the observations.

[33] See Section 5.2.2.

There is a clever way to get around this problem. We may be able to identify traits that are determined largely by a single locus. By repeating the procedure with different sets of parental lines which differ in various traits taken from within this special class of "single-locus" traits, we should be able to construct a genetic map for these loci, which we give the special name of **flag loci**.

Then, returning to the general case of traits that may often be determined by a great many loci we should be able to exploit the fact that any locus close to a flag locus can be putatively assigned a genotype by letting the trait value of the flag locus serve as a proxy. If the locus is so close to the flag locus that crossing-over never acts to separate the two, this "genotyping by proxy" is entirely warranted. As the distance to the nearest flag locus grows, the linkage assumption becomes more questionable, and some statistical modelling has to be done to recover information from the data.

We should ideally like to have at our disposal a set of flag loci that is not only large, but also distributed over the genetic map in an evenly spaced manner. Such an ideal **framework** minimises the errors we introduce with our trick of phenotyping by proxy. In the era of classical

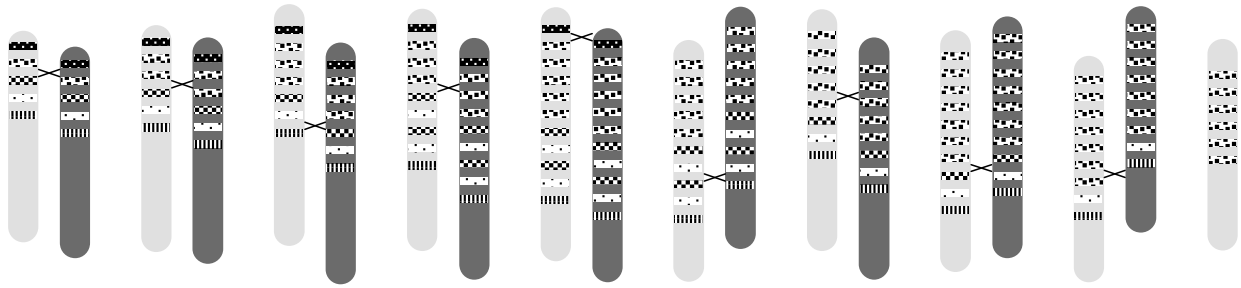


Fig. 7.7 A series of unequal crossing-overs generates a tandem repeat sequence. The original chromosome has five sequences that are sufficiently similar to give rise to mispairing. If this train of sequences is located in a non-coding region, the sequence will be selectively neutral, or virtually so. Thus, the cluster will wax and wane randomly, while it gradually becomes dominated by one of the five original sequences.

genetics, this desideratum posed a formidable challenge, since we are looking for loci with the flag property (i.e., being the sole or main determinant of the trait value) and the cost of observing the trait value should not be prohibitive. Attention was therefore focused on readily observable morphological traits; these represent the group of **classical** flag loci.

Another important group, besides the classical flag loci, is that of the **biochemical** flag loci. Here the differences between the gene products at a given flag locus are identifiable by biochemical means, e.g. staining, electrophoresis, or dedicated bioassays. Detection of this type of allelic variation is more involved than simply noting a gross morphological property such as flower colour.

The biochemical detection approach can be directed toward the DNA itself. Various features of the DNA sequence can be observed and quantitated, and whenever these differ between the parent lines, we have an **informative marker**. Since DNA markers can be located in the non-coding regions of the DNA, the term "flag locus" is too restrictive and the more general term "marker" is used; so that a flag locus is just one particular type of marker.

Satellite DNA constitutes an important class of DNA marker that occurs in non-coding regions (and therefore presumably has no or only very little influence on phenotype). At the molecular level, they consist of clusters of tandem repeats of short sequences of DNA.[34] These clusters are often highly variable between different inbred lines. The reason for this becomes apparent when we consider how they arise. The genesis of tandem repeat clusters is illustrated in Fig. 7.7. The driving mechanism is unequal crossing-over, shown in Fig. 2.21. Besides satellite DNA markers, there are various other classes which rely on different types of mutations, such as point mutations, insertions, and deletions.

[34]These repeated sequences can be as short as only a few bases or as long as over a hundred; the shortest type is also known as **microsatellite DNA** and the intermediate-length class is called **minisatellite DNA**.

7.2.1 Generalisation of the likelihood function

The general idea is to use marker information to guess each individual's genotype, in order to relate the latter to the observed trait value. More specifically, our aim is to formulate an probability distribution over the possible genotypes for each individual, based on the available marker information, which we denote as $\boldsymbol{M}_i$ for individual i. Given this information, the likelihood function with which we have been working thus far may be written in a more general form, as follows:

$$\mathbb{L}[\mu_{P_1}, \Delta\boldsymbol{\Upsilon}, \sigma^2] = \prod_{i=1}^{N} \sum_{\boldsymbol{\lambda}_{\female} \in \{0,1\}^n, \boldsymbol{\lambda}_{\male} \in \{0,1\}^n} \mathbb{P}[\boldsymbol{g}_i^{\female\male} = \boldsymbol{\lambda}_{\female} \text{ and } \boldsymbol{g}_i^{\male} = \boldsymbol{\lambda}_{\male} \mid \boldsymbol{M}_i] \times f(Y_i; \mu_{P_1} + \left(\boldsymbol{\lambda}_{\female}^{\mathsf{T}} + \boldsymbol{\lambda}_{\male}^{\mathsf{T}}\right) \cdot \Delta\boldsymbol{\Upsilon}, \sigma^2) \quad (7.13)$$

where N is the number of individuals, as before, n is the number of putative QTLs, $\boldsymbol{g}_i$ is a binary n-vector whose kth element equals 0 if, in the ith individual, the kth putative QTL derives from parental line 1 and 1 if the QTL is from parent 2. The vector $\Delta\boldsymbol{\Upsilon}$ contains the $\Delta\Upsilon$s at these putative loci, while terms of the form $(\boldsymbol{\lambda}^T \cdot \Delta\boldsymbol{\Upsilon})$ add up those $\Delta\Upsilon$s where the allele at the QTL is from parent 2. In the sum, $\boldsymbol{\lambda}$ is an index that ranges over all possible binary vectors of length n, i.e. the set $\{0,1\}^n$.

The novel element in the generalised expression is a conditional probability of the form

$$\mathbb{P}[\boldsymbol{g}_i = \boldsymbol{\lambda} \mid \boldsymbol{M}_i] \,,$$

which represents the probability that $\boldsymbol{g}_i$ equals a given value of λ, given the marker information $\boldsymbol{M}_i$. Applying Bayes' Rule we obtain:[35]

$$\mathbb{P}[\boldsymbol{g}_i = \boldsymbol{\lambda} \mid \boldsymbol{M}_i] = \mathbb{P}[\boldsymbol{g}_i = \boldsymbol{\lambda}] \frac{\mathbb{P}\left(\boldsymbol{M}_i \mid \boldsymbol{g}_i = \boldsymbol{\lambda}\right)}{\sum_{\boldsymbol{\lambda'}} \mathbb{P}\left[\boldsymbol{M}_i \mid \boldsymbol{g}_i = \boldsymbol{\lambda'}\right] \mathbb{P}[\boldsymbol{g}_i = \boldsymbol{\lambda'}]} \,. \quad (7.14)$$

This is useful provided that the probabilities on the right can be evaluated using a stochastic model of the genetic dynamics over the course of the particular breeding protocol followed in obtaining the data. It may not always be possible to obtain exact expressions for these probabilities, in which case we may have to resort to stochastic simulations (Monte Carlo simulation) or approximate the problem in a suitable way.

Another difficulty here is the assignation of putative QTL loci. One approach is to assume the QTL are near some informative marker. This is not unreasonable, provided that the analysis is interpreted with care: *if* a simple t-test yields a statistically detectable difference when the observed trait values are grouped according to a marker, this certainly points to a QTL in the vicinity of that marker. The converse is not quite as straightforward. Extending this approach to all available markers, we are led to analysis of variance (ANOVA);[36] in particular, ANOVA con-

[35] Check this; see Section 4.1 for Bayes' Rule.

[36] See Section 5.3.3.

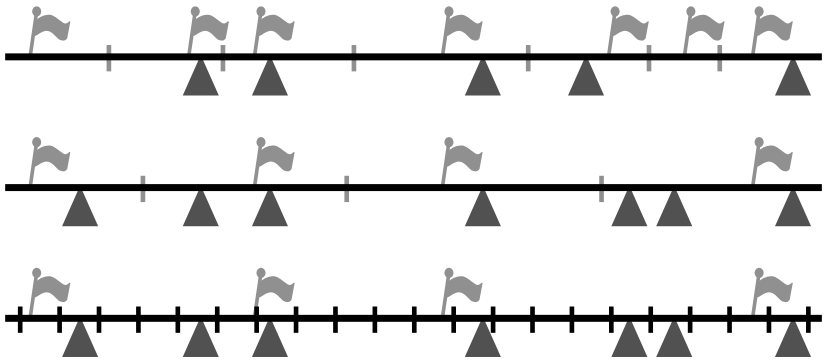


Fig. 7.8 The flags denote genetic markers scattered along the linear chromosome, whereas the triangles indicate the positions of QTLs that contribute to the trait of interest. Top: partitioning of the chromosome according to the markers (grey bars), typically containing at most one QTL and the simple approach is valid; middle: the markers are too scarce relative to the density of QTLs; bottom: a grid is imposed on the chromosome, and probabilistic "pseudo-marker" has to be devised for the grid points.

trasts represent a natural way to combine groupings by marker genotype into hypotheses regarding QTL effects for loci at these markers.[37]

The challenge posed by the general formalism of eqns (7.13) and (7.14) is formidable enough that a number of simplifying approaches have been put forward, a few of which for the subject of the next section.

7.2.2 Marker framework maps

A straightforward approach would be to associate each marker with a domain of the chromosome and treat each such domain as a locus; these domains may be defined as shown in Fig 7.8. Such a approach seems reasonable if these domains are roughly the size of actual gene loci: that is to say, the domains typically contain no more than one locus in the usual sense of the word.[38] In this case, the problem reduces to a regression problem. Otherwise, we can keep the problem in the regression form by means of a clever trick.

[38] Argue that the number of loci found in the domain associated with a given marker follows a Poisson distribution, and why explain why we must require that the parameter of this distribution is much smaller than 1.

Interval mapping

When it is reasonably safe to assume that the marker information $\boldsymbol{M}_i$ directly furnishes the genotype vectors $\boldsymbol{g}_i^{\varphi}$ and $\boldsymbol{g}_i^{\sigma}$, the likelihood function reduces to:

$$\mathbb{L}[\mu_{P_1}, \Delta\boldsymbol{\Upsilon}, \sigma^2] = \prod_{i=1}^{N} f(\mu_{P_1} + (\boldsymbol{g}_i^{\varphi} + \boldsymbol{g}_i^{\sigma}) \cdot \Delta\boldsymbol{\Upsilon}, \sigma^2) \,. \tag{7.15}$$

This can be recognised as the likelihood associated with the classical linear regression problem. To cast the problem in the traditional notation

[37] Unfortunately, there appears to be a tradition in the literature to emphasise the likelihood ratio relative to the all-but-uninteresting null hypothesis that *none* of the putative loci is a QTL, with displayed plots of the log of this ratio versus marker position, instead of similar plots with interval estimates for the $\Delta\boldsymbol{\Upsilon}$, which might often be more congenial to the underlying research objectives of the geneticist.

of linear regression, we define **genetic predictor** variables, as follows:

$$x_{ij} = \begin{cases} -1 & \text{if } i \text{ is homozygous for marker } j, \text{ stemming from } P_1 \\ 0 & \text{if individual } i \text{ is heterozygous for marker } j \\ +1 & \text{if } i \text{ is homozygous for marker } j, \text{ stemming from } P_2 \end{cases} \tag{7.16}$$

These predictors can be put to good use, as follows. Let

$$\Delta\Upsilon_j = \Upsilon(\alpha_j^{[2]}) - \Upsilon(\alpha_j^{[1]}) \tag{7.17}$$

where $\alpha_j^{[p]}$ is the allele at marker position j in parent p and let

$$\overline{\Upsilon}_j = \frac{\Upsilon(\alpha_j^{[1]}) + \Upsilon(\alpha_j^{[2]})}{2} \,. \tag{7.18}$$

Also, let M denote the number of markers. Then the additive genetic model takes on the following form:[39]

$$\begin{aligned} Y_i = \Upsilon_{A,i} + \Upsilon_{E,i} &= \sum_{j=1}^{M} \left(2\overline{\Upsilon}_j + x_{ij}\Delta\Upsilon_j\right) + \Upsilon_{E,i} \\ &= \overline{\overline{\Upsilon}} + \sum_{j=1}^{M} x_{ij}\Delta\Upsilon_j + \Upsilon_{E,i} \end{aligned} \tag{7.19}$$

where $\overline{\overline{\Upsilon}}$ is defined to be $2\sum_{j=1}^{M} \overline{\Upsilon}_j$. If we assume $\Upsilon_{E,i}$ to be an independently and normally distributed error term with mean zero, we can show that the maximum-likelihood estimator of the vector $\Delta\boldsymbol{\Upsilon}$ is obtained by minimising the quantity

$$S = \sum_{i=1}^{N} \left(Y_i - \overline{\overline{\Upsilon}} - \sum_{j=1}^{M} x_{ij}\Delta\Upsilon_j \right)^2 \tag{7.20}$$

with respect to the parameters $\overline{\overline{\Upsilon}}$ and $\Delta\Upsilon_1, \ldots, \Delta\Upsilon_M$.[40] In principle, this procedure establishes an **interval map** based on the marker-defined framework. This map can be visualised by plotting the estimated $\Delta\Upsilon_j$ as a function of j; this gives a profile along the chromosome indicating where the QTLs are located.

[39] Verify that we obtain the following: for $x_{ij} = -1$, $2\Upsilon(\alpha_j^{[1]})$; for $x_{ij} = 0$, $\Upsilon(\alpha_j^{[1]}) + \Upsilon(\alpha_j^{[2]})$; and $x_{ij} = 1$, $2\Upsilon(\alpha_j^{[2]})$.

[40] The quantity S is known as the **sum of squares**. Demonstrate that the parameter values obtained by minimising S are the MLEs, as claimed.

Grid mapping

The partitioning of the chromosome by marker-associated domains may be too coarse in the sense that some of the domains may contain quite a few classical loci. To deal with this complication, we lay down a grid over the chromosome that is fine enough to warrant the assumption that each grid point corresponds to at most one locus in the classical sense (see Fig. 7.8). For the sake of simplicity, we assume that the grid points are regularly spaced (in terms of centimorgans, not necessarily base-pair terms). Since the grid may locally be more finely spaced than the

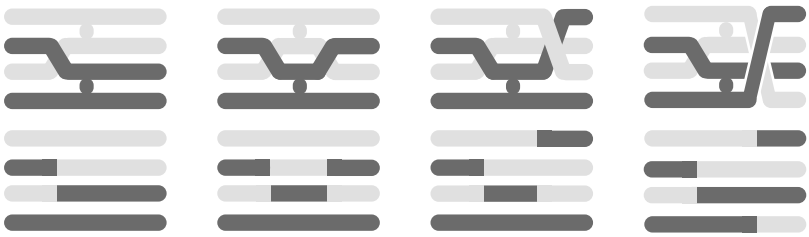


Fig. 7.9 A selection of possible crossing-over patterns, with the resulting chromosomes that segregate to the gametes. The simple pattern of the extreme left is the most common. All four sister chromatids in the tetrad can take part in crossings, and each can take part in one or more chiasmata.

marker-defined framework, there is a degree of uncertainty surrounding the homozygosity of the grid point q. Accordingly, we generalise the definition of the genetic predictor as follows:

$$x_{iq} = -\mathbb{P}_q^{\varphi}[1 \mid \boldsymbol{M}_i] + \mathbb{P}_q^{\sigma}[2 \mid \boldsymbol{M}_i] \tag{7.21}$$

where $\mathbb{P}_q^{\varphi}[p \mid \boldsymbol{M}_i]$ (respectively $\mathbb{P}_q^{\sigma}[p \mid \boldsymbol{M}_i]$) is the probability that the locus at grid point q on the maternal (respectively, patternal) chromosome has an allele from parent line p, given the marker information for individual i, denoted by $\boldsymbol{M}_i$. Apart from this generalisation,[41] the procedure remains basically the same, except that M in the definition of S, eqn (7.20), is replaced by Q, the number of grid points.[42]

The one remaining difficulty is the evaluation of the conditional probabilities given the marker information. In the back-cross pedigree we have been considering, one of the chromosomes is known to contain markers only from one the inbred, homozygous, parent lines, and therefore one of the two terms in eqn (7.21) can be evaluated immediately to be 0 or 1. For the other chromosome, we have to take into account the possible crossing-over events taking place between the parental chromosomes in the generation of the F_1 gametes (which gave rise to the F_2 offspring on which we assume the observations have been performed), as indicated in Fig. 7.3. Inasmuch as multiple crossing-over events are possible (Fig. 7.9), the situation has to be analysed with some care.

Let ν_q denote the distance (in cM) between the nearest informative markers flanking grid point q. We assume that n_c, the number of crossovers between two these to markers, follows a Poisson distribution with parameter μ_q. Furthermore, let $\lambda_q \in [0, 1]$ be such that the distance between grid point q and the nearest upstream marker is $\lambda_q\nu_q$ so that the distance between q and the nearest downstream marker is $(1-\lambda_q)\nu_q$; see Fig. 7.10. From these definitions, it follows that the number of those upstream of grid point q follows a binomial distribution with parameters n_c and λ_q.

The markers may agree on which parental chromosome they derive

[41] Check that x_{iq} as defined by eqn (7.21) takes values in the interval $[-1, 1]$ and agrees with the earlier, "deterministic" definition of eqn (7.16) when the probabilities take on extreme values. This verifies that the generalisation subsumes the simpler case of a sufficiently dense marker framework.

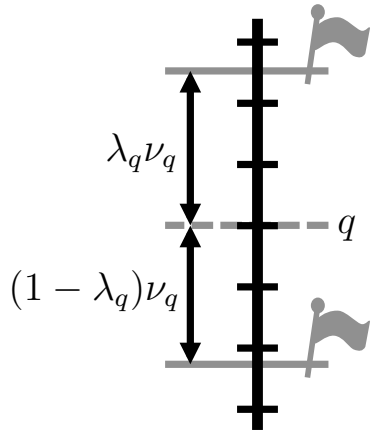


Fig. 7.10 Grid point q situated at relative position λ between two markers separated by a distance ν_q. Upstream is upwards.

[42] The use of continuous genetic predictor variables on the interval $[-1, +1]$ amounts to a slight sleight of hand: comparison of equations (7.13) and (7.19) reveals that we are in effect substituting a pdf of the form $f(Y; \sum_k \mu_k, \sigma^2)$ for what, strictly speaking, is a mixed distribution of the form $\sum_k f(Y; \mu_k, \sigma^2)$ which has the same mean but a somewhat more complex shape; the gain is in computational speed.

from, or they may disagree. The probability of the latter is given by

$$P_{\text{even}} = \sum_{k=0}^{\infty} \frac{\mu_q^{2k} \exp\{-\mu_q\}}{(2k)!} = \frac{1}{2}\left(1 + \exp\{-2\mu_q\}\right) \tag{7.22}$$

since an even number of cross-overs has no net effect on the provenance agreement of the flanking markers.[43] On the other hand, the flanking markers will derive from different parental lines with the complementry probability $P_{\text{odd}} = 1 - P_{\text{even}}$.[44]

Given that the flanking markers agree, the probability that n_c cross-overs have occurred is

$$P(n_c) = \frac{\mu_q^{n_c} \exp\{-\mu_q\}}{n_c! P_{\text{even}}}$$

for $n_c = 0, 2, 4, \ldots$ and zero otherwise. If the flanking markers correspond to different parents, the probability that n_c cross-overs have occurred is

$$P(n_c) = \frac{\mu_q^{n_c} \exp\{-\mu_q\}}{n_c! P_{\text{odd}}}$$

for $n_c = 1, 3, 5 \ldots$ and zero otherwise. Finally, let ℓ denote the parent line from which the upstream marker derives. The grid point at q will agree if there has been an even number of cross-overs between q and the upstream marker. The probability that this will be the case is given by the following expression:

$$\mathbb{P}_q[\ell \mid \boldsymbol{M}_i] = \sum_{n_c=0}^{\infty} P(n_c) \sum_{s=0}^{\text{ent}(n_c/2)} \binom{n_c}{2s} \lambda_q^{2s} (1-\lambda_q)^{n_c - 2s} \tag{7.23}$$

where $\text{ent}(n_c/2)$ denotes the largest integer smaller than $n_c/2$; the sum with index s is over even numbers of cross-overs upstream of locus q.

7.2.3 The number of quantitative trait loci

Three interrelated questions about QTLs were posed at the beginning of this chapter: where are they, how large are their effects, and how many are there for a given trait of interest? The first two questions we have answered, and we have seen that the answer takes the form of a estimate of $\Delta\Upsilon$ as a function of grid position. The locations of the *major* contributing loci (or grid positions) and the prevailing recombination rates between them are perhaps most important, both from the point of view of evolutionary dynamics and of practical genetics (e.g., crop improvement), but the binary question of whether or not any given position *does* or *does not* contribute imposes itself. In a sense, this is not the right question to ask: attempts to answer it lead to a perplexities and the answer need not even be all that informative.

We consider here an argument in favour of the viewpoint that all positions contribute. Furthermore, we demonstrate that this perspective is entirely in keeping with the representation of the additive genetic component by a Gaussian term.

[43] To derive the last equality, write down the Taylor series for e^{μ} and $e^{-\mu}$, then form their sum and rearrange.

[44] By definition $1 - \nu_q = P_{\text{even}}$ and therefore $\mu_q \approx \nu_q$ when $\nu_q \ll 1$, a result that can be checked by considering the Taylor expansion of the exponential in eqn (7.22).

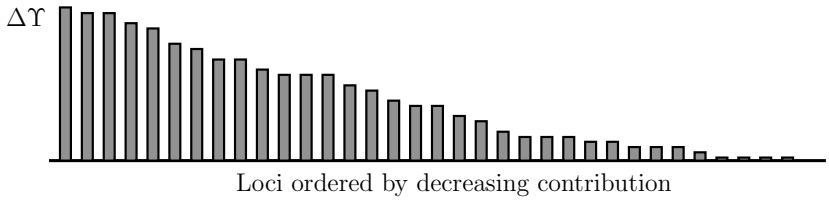


Fig. 7.11 The contributions made by the loci taper off into less and less importance ones.

A statistical quandary

Inferential statistics can be used to determine whether the $\Delta\Upsilon$ at a given position is significantly different from zero. One problem is whether the estimate is biased: if the experiment where repeated many times, would the average of the estimated values converge to the underlying "true" parameter value? Even if the estimated $\Delta\Upsilon$ is unbiased, a problem remains, because among the loci that are on one particular occasion detected by a statistical procedure (i.e., are deemed "statistically significant"), the estimated values are more likely to have erred upwards than downwards on that particular occasion. This effect is most pronounced for small effects. The converse bias affects the loci that have been rejected as "non-significant"; their estimated values are more likely to have erred downwards. This bias is not inherent in the estimation procedure, but results from the desire to divide loci into a contributing and a non-contributing class.[45]

[45]In other words, the error is engendered by an insistence on a *yes/no*-type answer when it would be more appropriate to ask for a *value* (and then perhaps worry about its accuracy).

Accept that all loci contribute

Inasmuch as the perplexities flow from the preconditioning on statistical detectability, the way out seems obvious: just refrain from this preconditioning.[46] This amounts to accepting that all positions contribute, in most cases a minute amount that is as close to zero as to make no difference. The picture that accompanies this point of view is shown in Fig. 7.11: the loci have been ordered according to descending (estimated) $\Delta\Upsilon$. In practice, we would have to truncate this graph at some point and neglect the contributions made by the long tail to the right of this point; but where this point is to be chosen is just a pragmatic issue.

The QTL distribution for a large number of loci

Consider the contribution from one of the two parents:

$$\Upsilon = \sum_{\lambda \in \mathbb{G}} \Upsilon_\lambda$$

where we have dropped the male or female symbol. On the point of view that all loci contribute in principle, we should allow $\mathbb{G}$ to cover the entire

[46]The situation is somewhat different for the estimated dominance deviations. Here statistical testing at individual loci has a good justification, since a strong rejection of the additive null hypothesis points to an interesting mechanism underpinning the observed epistasis. However, for reporting purposes, it is good practice to present a trace of ε as a function of grid index q along with $\Delta\Upsilon$.

genome in principle. Assume for the sake of simplicity that the loci can be divided into n unlinked clusters of tightly linked loci. Clearly $n \leq \mathbb{G}$, but we should still expect that $n \gg 1$. The limit $n \to \infty$ is therefore relevant. To avoid the need for additional and cumbersome notation, we shall label the clusters by the locus index λ, remembering that each such "locus" is in fact a cluster which possibly contains a single locus, but might just well comprise very many loci in the usual sense of the word.

Suppose that at each locus, there are G_λ different possible values for Υ_λ.[47] The number G_λ may vary greatly with λ, since some of the loci may comprise few actual loci and others may be large but tightly linked clusters. Denote the values as $\Upsilon_{\lambda,i}$ for $i = 1, \ldots, G_\lambda$. Thus, for each λ we have a separate distribution, at the gamete population level, defining the probabilities of encountering the values, i.e.

$$p_{\lambda,i} = \mathbb{P}[\Upsilon_\lambda = \Upsilon_{\lambda,i}] \, .$$

[47] We are about to derive the Central Limit Theorem for this particular setting. The motivation for outlining the argument in some detail is that, at least at the undergraduate level, the proof given is usually restricted for *identically* distributed variables, which is not the case here.

Let μ_λ and σ_λ^2 denote the mean and variance of this distribution. The moment-generating function[48] of $\Upsilon_\lambda - \mu_\lambda$ is given by

[48] See Exercise 4.13.

$$m_\lambda(t) = \sum_{i=1}^{G_\lambda} \exp\{t\Upsilon_{\lambda,i}\} p_{\lambda,i} \tag{7.24}$$

which by Taylor's theorem can be written as follows:

$$m_\lambda(t) = 1 + \frac{\sigma_\lambda^2 t^2}{2} + \frac{\left(m_\lambda''(\xi) - \sigma_\lambda^2\right) t^2}{2} \tag{7.25}$$

where we have used the properties $m_\lambda(0) = 1$ and $m_\lambda'(0) = 0$, and ξ is an unknown value between 0 and t. Define the overall variance as a straightforward average:

$$\sigma^2 = \frac{1}{n} \sum_{\lambda=1}^{n} \sigma_\lambda^2 \, . \tag{7.26}$$

We now show that the normalised variate $Z_n = \sum_{\lambda=1}^{n} (\Upsilon_\lambda - \mu_\lambda)/(\sqrt{n}\,\sigma)$ follows the standard normal distribution. Consider the natural logarithm of the moment-generating distribution of Z_n:

$$\begin{aligned} \ln M_{Z_n}(t) &= \sum_{\lambda=1}^{n} \ln m_\lambda(t/(\sqrt{n}\,\sigma)) \\ &= \sum_{\lambda=1}^{n} \ln \left(1 + \frac{\sigma_\lambda^2 t^2}{2n\sigma^2} + \frac{\left(m_\lambda''(\xi) - \sigma_\lambda^2\right) t^2}{2n\sigma^2} \right) \end{aligned} \tag{7.27}$$

where the second equality is due to eqn (7.25). Since $|\xi|$ is bounded above by $|t|/(\sqrt{n}\,\sigma)$, it follows that for fixed t, $\xi \to 0$ as $n \to \infty$. But

$m''_\lambda(0) = \sigma^2_\lambda$, and therefore $m''_\lambda(\xi) - \sigma^2_\lambda \to 0$ as $n \to \infty$. We thus have

$$\lim_{n\to\infty} \ln M_{Z_n}(t) = \sum_{\lambda=1}^{n} \lim_{n\to\infty} \ln\left(1 + \frac{\sigma^2_\lambda t^2}{2n\sigma^2} + \frac{\left(m''_\lambda(\xi) - \sigma^2_\lambda\right) t^2}{2n\sigma^2}\right)$$
$$= \frac{t^2}{2} \sum_{\lambda=1}^{n} \frac{(\sigma_\lambda/\sigma)^2}{n} = \frac{t^2}{2} \tag{7.28}$$

which completes the proof, since the moment-generating function of the standard normal distribution is $\exp\{t^2/2\}$. Thus for large n, it is reasonable to treat Υ as a Gaussian variate with mean $\sum_{\lambda=1}^{n} \mu_\lambda$ and variance σ^2 as defined in eqn (7.26).

7.3 Multi-environment QTL: reaction norms

7.4 Evolutionary consequences

7.4.1 Linkage of co-adaptations

Further reading

- J. Maynard Smith (1989). *Evolutionary Genetics*. Oxford University Press, Oxford.

Exercises

(7.1) Suppose that the genome of an organism of interest has 10,000 loci.
(a) Assuming that crossing over is so vigourous that the loci are effectively unlinked, derive a formula for the number of loci that fail to be homozygous after n generations of inbreeding.
(b) Explain why the assumption of part (a) represents a "worst-case" scenario.
(c) How many generations of inbreeding are required to reduce the probability that one or more loci fail to be homozygous to less than one percent?

(7.2) Given that there are m alleles that can appear at a given locus, state the probability that the initial breeding pair has $1, 2, \ldots, m$ alleles between them.

(7.3) A desideratum is that all allelic variants that occur within a polymorphism are represented by the repertoire of inbred lines. For a locus that has a polymorphism of m alleles, state how many inbred lines must be established and maintained in order to reduce the probability that one or more alleles are nor represented to less than one percent.

(7.4) The following diagram is a genetic map of a hypo-

thetical chromosome (genes are indicated by capital letters):

```
                  KM
A B C   D   E F G   H   I JLN O P Q   R   S T U
```

By contrast, the following diagram is a physical map of the same chromosome:

```
A B C D E F G H I J K L M N O P Q R S T U
```

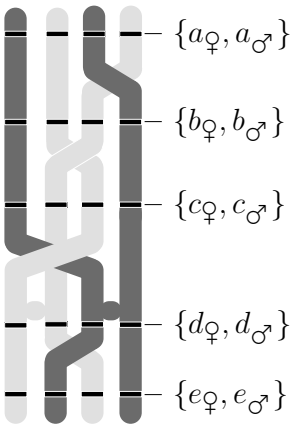
Indicate where the hot spots are located. Also indicate the probable position of the centromere.

(7.5) The table records the distances (in cM) between 11 loci on a hypothetical linear chromosome. From these data, reconstruct a genetic map of the chromosome.

(7.6) The table records the distances (in cM) between 6 loci on a hypothetical linear chromosome; with missing data indicated by blanks. From these data, reconstruct a genetic map of the chromosome.

(7.7) (a) Sketch and label the counterpart of Fig. 7.6 for the case of two loci. (b) Explain the symmetry apparent in Fig. 7.6; can this phenomenon be exploited to improve the analysis of the data?

(7.8) After DNA replication, each chromosome becomes a pair identical sister chromatids; homologous pairs of chromosomes thus form a **tetrad** of four chromatids. Within the tetrad, crossing over between chromatid strands in all possible combinations may occur, as shown in the diagram below.



Consider five loci, indicated by the letters *a*–*e*, at which the maternal (dark grey) chromatids bear the allele indexed with ♀ and the paternal (light grey) chromatids bear the allele indexed with ♂.

(a) Given the cross-overs indicated in the diagram, state the genotypes of four cells haploid cells produced after meiosis completes.

(b) In general, how many different recombinant genotypes are possible for the haploid daughter cells? How many of these are recombinant?

(c) There is a cross-over event between the levels of the b and c loci, but no recombination occurs between these two loci. Explain why not.

(d) Considering only two loci at a time, there are four haploid genotypes, out of which two are recombinants. On the basis of the cross-overs indicated in the diagram, state how many recombinant chromosomes are obtained for the allele pairs (a, b), (b, c), (c, d), (d, e), and (a, e). What is noticeable about these numbers, and how can this be explained?

(7.9) Consider a chromatid tetrad and two alleles at the very ends of the chromosomes, e.g., alleles a and e in the diagram of Exercise 7.8. If these alleles were on loci on separate loci, the probability that a gamete receives the maternal allele of the one locus and the paternal allele of the other equals $\frac{1}{2}$; this is the maximum recombination fraction. On the other hand, multiple crossings theoretically allow a *higher* recombinant fraction. Could alleles linked by their location on the same chromosome have a recombinant fraction exceeding $\frac{1}{2}$? To address this question, imagine travelling from one end of the tetrad to the other, keeping track of the number of recombinant haploid genotypes that would result if no more crossings further down the chromosome were to be observed.

(a) Explain why this number must be 0, 2, or 4.

(b) Describe an example of a crossing-over that does not change the number.

(c) We restrict the remainder of the analysis to cross-over events that change the number. Given that such an event occurs explain the following probabilities $\mathbb{P}[y \mid x]$ where x is the number just prior to (i.e., upstream from) the event and y is the number immediately after it:

$$\mathbb{P}[2 \mid 0] = \mathbb{P}[2 \mid 4] = 1 \; ;$$
$$\mathbb{P}[0 \mid 2] = \mathbb{P}[4 \mid 2] = \frac{1}{2} \; ;$$
$$\mathbb{P}[0 \mid 4] = \mathbb{P}[4 \mid 0] = 0 \; .$$

(d) Let $p_{x,i}$ denote the probability that the number has the value x after i events and collect these probabilities in a vector $\boldsymbol{p}_i$. The following relationship holds:

$$\boldsymbol{p}_{i+1} = \boldsymbol{M} \cdot \boldsymbol{p}_i \tag{7.29}$$

where $\boldsymbol{M}$ is a 3×3 matrix. Write down this matrix in full, displaying the nine elements explicitly.

(e) Given that $\boldsymbol{p}_0 = [1, 0, 0]^{\mathsf{T}}$, calculate $\boldsymbol{p}_i$ for

$i = 1, 2, 3, 4, 5, 6, 7, 8$.

(f) If the number of recombinant-number altering events along the length of the tetrad follows a Poisson distribution with mean μ, the probability that the number of events is event is given by:

$$P_{\text{even}} = \sum_{k=0}^{\infty} \frac{\mu^{2k} e^{-\mu}}{(2k)!} \ . \tag{7.30}$$

Show that

$$P_{\text{even}} = \frac{1}{2} \left(1 + \exp\{-2\mu\}\right) \ .$$

(g) Let $P_{\text{odd}} = 1 - P_{\text{even}}$. Show that

$$P_{\text{odd}} = \frac{1}{2} \left(1 - \exp\{-2\mu\}\right) \ .$$

(h) Give a formula for the probability that the number of events is even *and* greater than 1.

(i) Let R be the number of recombinants at the end of the imaginary journey, i.e., the actual number of recombinants, with $R \in \{0, 2, 4\}$. Give expressions for $\mathbb{P}[R = 0]$, $\mathbb{P}[R = 2]$, and $\mathbb{P}[R = 4]$.

(j) Show that[49]

$$\mathbb{E}[R/4] = \frac{1}{2} \left(1 - e^{-\mu}\right) \ . \tag{7.31}$$

(k) Show that $\mathbb{E}[R/4] \approx \mu/2$ when $\mu \ll 1$.

(l) If each cross-over event were to leave a "scar" on the chromatid, what would be the mean and variance of the distribution of the number of "scars"?

(7.10) Prove and explain why tandem repeat clusters eventually tend to be dominated by repeats of one particular sequence, as illustrated in Fig. 7.7.

Table for Exercise 7.5

	A	D	G	J	L	M	O	P	U	Y	Z
A	0.	1.16206	2.67695	3.05409	0.388988	2.70874	1.1782	0.94955	2.72386	0.757281	0.400503
D	1.16206	0.	1.51489	4.21615	1.55105	3.8708	0.0161355	2.11161	3.88592	0.404781	0.761559
G	2.67695	1.51489	0.	5.73104	3.06594	5.38569	1.49875	3.6265	5.40081	1.91967	2.27645
J	3.05409	4.21615	5.73104	0.	2.6651	0.345353	4.23229	2.10454	0.33023	3.81137	3.45459
L	0.388988	1.55105	3.06594	2.6651	0.	2.31975	1.56719	0.560562	2.33487	1.14627	0.789492
M	2.70874	3.8708	5.38569	0.345353	2.31975	0.	3.88694	1.75919	0.0151227	3.46602	3.10924
O	1.1782	0.0161355	1.49875	4.23229	1.56719	3.88694	0.	2.12775	3.90206	0.420917	0.777694
P	0.94955	2.11161	3.6265	2.10454	0.560562	1.75919	2.12775	0.	1.77431	1.70683	1.35005
U	2.72386	3.88592	5.40081	0.33023	2.33487	0.0151227	3.90206	1.77431	0.	3.48114	3.12436
Y	0.757281	0.404781	1.91967	3.81137	1.14627	3.46602	0.420917	1.70683	3.48114	0.	0.356778
Z	0.400503	0.761559	2.27645	3.45459	0.789492	3.10924	0.777694	1.35005	3.12436	0.356778	0.

Table for Exercise 7.6

	A	E	G	I	M	N
A			0.94955		0.388988	
E	3.05409				2.6651	
G				1.75919		1.77431
I		0.345353				0.0151227
M			0.560562	2.31975		
N	2.72386	0.33023				

[49] As for the question posed at the start of the exercise, the answer found on the present assumptions is no, as can be seen from eqn (7.31), but it should be noted that the theoretical possibility of positive cooperatively among chromatids crossings within a tetrad has not been explored here.