17

Principles of Marker-based Mapping

Version 8 Dec. 2022

In a classical Mendelian analysis, scorable genetic variants are used to map and characterize genes. This analysis requires that individual phenotypes are *highly informative* about their underlying genotypes. Quantitative genetics (QG) deals with the opposite setting, wherein the phenotype provides *very little information* on the underlying genotype. Because of this distinction, analysis in the classical QG setting is based on genetic variances rather than on the underlying genes themselves (Chapters 4–7). While this variance-based approach is sufficient in many situations, we ultimately would like to move away from this strictly statistical framework towards direct examination of the effects of the individual **quantitative-trait loci** (**QTLs**) that underlie trait variation.

As discussed below, there are occasions when one can choose suitable **candidate loci** based on biological knowledge of the character, but in most situations this is not possible. Fortunately, QTLs can be assayed *indirectly* by using linked **marker loci**. This indirect approach had long been recognized (e.g., Payne 1918; Sax 1923), but until the 1980s was regarded as of minor importance because of the lack of a sufficient number of genetic markers to make it a viable method. In the genomics era, this situation changed dramatically, to the point where now a whole genome sequence for any organism can now be obtained with not much more effort than was required in the past to score a few dozen molecular markers. As mentioned in Chapter 1, a direct consequence of this new ability to have an essentially endless supply of markers for any organism of interest resulted in a dramatic paradigm shift in the field. Historically, quantitative genetics was largely based on statistical inference of phenotypes of interest by comparing individuals with known degrees of relationship (i.e., the coefficient of coancestry Θ; Equation 7.2a). Modern quantitative genetics uses molecular marker information to augment, or even replace, information on known relationships (Chapter 8).

There are two different, but not exclusive, board strategies that use marker information. First, as discussed in Chapter 8, markers allow for a quantitative-genetic analysis when relationships among a set of individuals are either ambiguous or even unknown (as is often the case in most natural population settings). Here, marker information is used to estimate the degree of relationships, but otherwise the machinery of quantitative genetics remains the same. Second, and the focus of the next several chapters, is to (largely) ignore relationships and instead search for marker-trait correlations, either using linkage following a cross or within a pedigree, or using linkage disequilibrium (LD) within a population.

This chapter introduces the basics of marker-based mapping, while the next several chapters consider specific refinements for linkage-based analysis using inbred lines (Chapter 18) and outbred pedigrees (Chapter 19), and linkage-disequilibrium analysis in a population (Chapter 20). Our discussion in this chapter is loosely organized by historical developments of marker-based methods. We start by considering classical approaches for quantifying the effects of entire chromosomes (or chromosomal segments). We then turn to finer genetic resolution, examining a variety of topics concerning markers and genetic maps, including the use of markers to facilitate construction of nearly isogenic lines, construction of specialized mapping populations for linkage-based mapping, and the use of population samples for LD-based mapping, concluding by discussing tests for candidate loci.

CLASSICAL APPROACHES

In some species, the clever use of special biological features (e.g., balancer chromosomes and the lack of male recombination in *Drosophila*) allows one to assay the effects of whole chromosomes on characters of interest. Because such **chromosomal assays** can be done with just a few morphological markers, they were the main method of marker-based trait analysis before the advent of molecular markers. Although restricted to a few species, such assays nonetheless have provided insight into the genetic architecture of a number of characters.

Chromosomal Assays

The idea here is straightforward: using genetic markers, a chromosome (or chromosomal segment) from one line is substituted into an otherwise standard genetic background (usually an inbred line to minimize background variance). A collection of lines in which each chromosome is changed against an otherwise standard background—and hence the ability of screen every individual chromosome—is referred to as a set of **chromosome substitution strains**, or **CSSs** (Nadeau et al. 2000).

Drosophila is highly amenable to such studies given its short generation time, few chromosomes, a well-characterized genetic map, the existence of balancer strains (e.g., Figure 5.7), and the lack of recombination in males. Application of these features allows intact chromosomes from one line to be substituted into another line. Chromosome substitution can also be done in wheat using certain genetic tricks that allow construction of individuals carrying an extra chromosome (e.g., Sears 1953; Law 1966; Snape et al. 1977; Law and Gale 1979; Choo 1983). The effects of large chromosomal segments have also been examined in maize through the use of reciprocal translocations (e.g., D. Robertson 1989) and in mice by using marked chromosomes (e.g., Kluge and Geldermann 1982). Finally, as reviewed by Nadeau et al. (2000) and Shao et al (2008), collections of CSSs have been constructed for probing mouse- and rat-based models of human disease.

In such investigations, given that each chromosome or chromosomal segment typically contains a significant fraction of the total genome, the considerations discussed in Chapter 11 apply, and it is best to speak of ***genetic factors***, rather than *individual genes*. In addition to providing a very coarse estimate of the location of such factors, chromosomal assays provide estimates of the *minimum number* of underlying loci, the degree of dominance measured by the composite (aggregate) effects of loci on the chromosome or chromosomal segment, and the degree of epistasis as measured by interactions between assayed chromosomal units.

With sufficiently large sample sizes, even factors of very small effect can be detected with this approach, and the types of characters that can be analyzed are limited only by the imagination of the investigator. For example, in *D. melanogaster*, Caligari and Mather (1988) used chromosomal assays to localize factors for aggression and competitive response to chromosomes 2 and 3. Sokolowski (1980) examined larval foraging behavior, mapping genetic factors for larval locomotor behavior to chromosome 2 and for feeding rates to both chromosomes 2 and 3. Luckinbill et al. (1988) localized factors deferring senescence to all major chromosomes, with chromosome 3 accounting for roughly 70% of the observed variation in females. Shao et al (2008) reviewed the results from 90 blood, bone, and metabolic traits in a mouse CSS panel and 54 additional traits in a rat CSS panel, finding that traits tended to be highly polygenic, with strong episatic interactions between chromosomes.

Estimating the additive, dominance, and epistasic composite effects associated with particular chromosomes is straightforward. For example, let X_B and X_b represent the X chromosomes from two different inbred strains. The composite amount of dominance is estimated by comparing the mean phenotypic values of X_BX_B, X_BX_b, and X_bX_b females in a standardized genetic background. Nonadditive interactions (epistasis) between chromosomes can be detected in a similar manner by considering pairs of chromosomes (Robertson and Reeve 1953; Robertson 1954; Cooke and Mather 1962). Using these approaches, the genetic architectures for a number of characters has been examined in *Drosophila* (Table 17.1). Characters correlated with fitness tend to show epistasis and directional dominance, while those presumably weakly correlated with fitness do not. A serious limitation of these assays is the size of the unit of analysis—a chromosome or large chromosomal segment

Table 17.1 The genetic architecture of various characters in *Drosophila* inferred by chromosomal assay. Unless otherwise indicated, all analyses are for *D. melanogaster*. Most of these results were obtained by comparing differences between selected lines. Results for many characters were obtained by both whole and segmental chromosomal assays. Directional dominance occurs when multiple assayed regions show the same direction of dominance (e.g., high values tending to be dominant over low values). Features not examined are denoted by a dash. (Expanded from a shorter table in Kearsey and Kojima 1967.)

	Dominance			
Character	Present	Directional	Epistasis	Reference
Fitness-related Characters				
Viability	Yes	Yes	Yes	Breese and Mather 1960
	—	—	Yes	Seager and Ayala 1982
	No	—	Weak	Ferrari 1987
Male mating activity	—	—	No	Kosuda 1993
Egg hatchability	Yes	Yes	Yes	Kearsey and Kojima 1967
Fecundity	Yes	Yes	—	Keller and Mitchell 1964
Egg-pupal survival	Yes	Yes	—	Keller and Mitchell 1964
Development time	Yes	Yes	—	Keller and Mitchell 1964
	No	—	Weak	Ferrari 1987
Progeny yield	Yes	Yes	Yes	Barnes 1966
Egg production				
D. melanogaster	Yes	Yes	—	Keller and Mitchell 1964
D. pseudoobscura	Yes	—	Yes	Kojima and Kelleher 1963
Physiological Characters				
DDT-resistance	Yes	No	Yes	Dapkus and Merrell 1977
ADH activity	Yes	Yes	Yes	McDonald and Ayala 1978
G6PD activity	—	—	Yes	Miyashita and Laurie-Ahlberg 1984
6PGD activity	—	—	Yes	
Morphological Characters				
Abdominal bristles	Weak	No	Weak	Keller and Mitchell 1962
	Weak	No	Weak	Breese and Mather 1957
Sternopleural bristles	Weak	No	Weak	Breese and Mather 1957
	Weak	No	No	J. Hill 1964
Thorax length	Yes	Yes	Yes	Robertson 1954
	Weak	No	Weak	Keller and Mitchell 1962
Wing length	Yes	Yes	Yes	Robertson 1954
	No	—	No	Keller and Mitchell 1962
Body weight				
D. melanogaster	No	—	Weak	Kearsey and Kojima 1967
D. pseudoobscura	Weak	No	Weak	Frahm and Kojima 1966

can contain QTLs having effects in opposite directions, canceling the individual effects and leaving a composite effect close to zero. This is especially true of epistatic effects (Hastings 1986).

Natural chromosomal assays become possible when chromosomal rearrangements are segregating within a population. Because chromosomal inversions suppress recombination, alternative inversions can be treated as intact, nonrecombining units. This approach was used extensively by Dobzhansky in his studies of the fitnesses of natural inversions in *Drosophila* (Lewontin et al. 1981), and has also been applied to morphological characters. For example, Hasson et al. (1992) found that three inversions on chromosome 2 accounted for at least 25% of the total variance for thorax length in *Drosophila buzzatii* (also see Ruiz et al. 1991). Two of the inversions showed dominance, while a third showed apparent (potentially associative) overdominance. Other studies have detected segregating rearrangements with effects on body size in the grasshopper *Moraba scurra* (White and Andrew 1960), on wing length in the seaweed fly *Coelopa frigida* (Butlin et al. 1982; Wilcockson et al. 1995), and on wing length in *D. melanogaster* (Herández et al. 1993).

Thoday's Method

Thoday (1961, 1979) introduced a considerable refinement for crude chromosomal assays. His method is of substantial interest for purely historical reasons, but it also points out potential limitations of the flanking-marker mapping methods discussed in subsequent chapters. Consider a hypothetical QTL between two linked markers, with allele *Q* fixed in one inbred line (the **tester line**, also fixed for alleles *A* and *B* at the flanking marker loci), and *q* in another line (fixed for markers *a* and *b*). Substituting an *aqb* chromosome into the tester line gives *AQB/aqb* heterozygotes, and these individuals are subsequently backcrossed to the tester line. Recombinant *aB* progeny are then scored for the trait of interest. Since these recombinant chromosomes are either *aqB* or *aQB*, a QTL with detectable effects is indicated by the presence of two distinct character classes (corresponding to *QQ* and *Qq* individuals). The presence of a linked QTL outside of the interval can also generate distinct classes, but this is not an issue if only one recombination event occurs per chromosome. The main limitation of Thoday's method is the difficulty of isolating single recombinant chromosomes and propagating them intact (without further recombination) in organisms other than *Drosophila*.

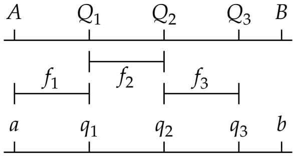
Example 17.1. A large sample of *Ab* recombinant chromosomes was generated from a cross between an *ab* line with negative character value and an *AB* tester line with a mean of zero. Each recombinant chromosome was used to create a line with a recombinant *Ab* chromosome and an *AB* tester chromosome in an otherwise genetically homogeneous background. Construction of such lines requires the use of balancer chromosomes, which prevent any further recombination from occurring (see Figure 5.7). The mean character values from the resulting lines formed two distinct clusters, one (comprising 43%) clustering around -1.5, the other clustering around 0. Hence, there appears to be a single factor (at our level of resolution), with the 0 mean class corresponding to Q/Q and the -1.5 class corresponding to Q/q, implying $\mu_{QQ} - \mu_{Qq} = 1.5$. The frequencies of these classes among *Ab* recombinant chromosomes imply that the *A–Q* recombination fraction is 43% of the *A–B* recombination fraction (assuming no interference).

McMillan and Robertson (1974) found that the power of Thoday's method is generally very poor. To achieve high power, both the number of recombinant chromosomes (N) and the number of replicates (n) per recombinant chromosome need to be large. Even if a factor has a large effect, it will remain undetected unless the sample of recombinant chromosomes contains both factor alleles, which requires a sufficiently large N. To see this result, set the distance between the two markers to one, and let f ($0 \leq f \leq 1$) denote the relative distance between factor and marker *A*, so that a fraction f of the *aB* recombinants are *aqB* and $1 - f$ are *aQB*. The probability that N recombinant (*aB*) chromosomes all contain the same QTL allele is $f^N + (1-f)^N$. If f is close to zero or one, N has to be considerable for a sample to contain a sufficient number of both QTL genotypes for detection. The power for detection is maximized when the factor is exactly between the two markers (leading to expected equal numbers of aqB and aQB). Power graphs are given by McMillan and Robertson (1974).

Thoday's method can be extended to allow for multiple factors between two markers, but not without complications. As shown in Example 17.2, the estimation of map distances and effects of multiple QTLs between markers *critically* assumes that all QTL alleles in the line being examined have *the same directional effect* (McMillan and Robertson 1974). This caveat will be a recurring theme for molecular-marker approaches: ***these methods can be compromised if the line being examined contains linked QTL alleles of opposite effects***.

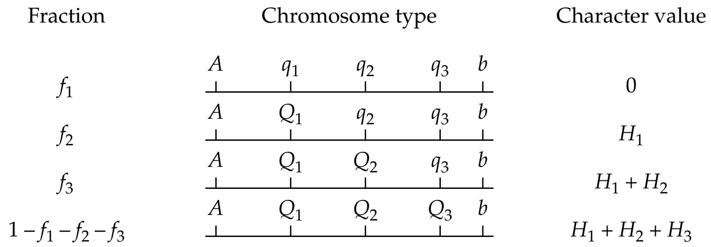
Data from inbred lines show that such situations are not uncommon (Chapter 18).

Example 17.2. Suppose there are three QTLs between the marker loci *A* and *B*, with

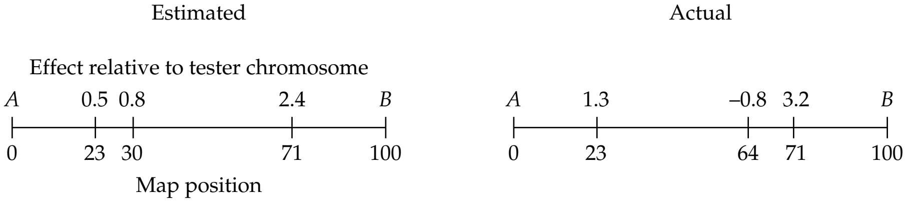


Assuming no double crossovers, let f_1, f_2, f_3, and $1 - (f_1 + f_2 + f_3)$, respectively, be the fractions of recombination between flanking markers that occur between *A* and the first QTL, between the first and second QTL, the second and third, and the third and *B*. (As an aside, for detailed power calculations, one can use these frequencies as the success probabilities in a multinomial distribution. Equation 2.20a then returns the probability of obtaining a particular configuration of these three classes given N recombinant chromosomes.) Let Q_i denote alleles from the high line, which have effect $H_i > 0$ relative to alleles q_i from the low line.

The expected frequencies and effects of *Ab* recombinant chromosomes are obtained as follows. As shown in the following figure, a fraction f_1 of *Ab* chromosomes have no high-line alleles, as the crossover occurred between *A* and Q_1. Likewise, a fraction f_2 contain high-line allele Q_1 with effect H_1. The last two classes follow similarly. Thoday's method estimates the H_i and f_i by making the *assumption* that all alleles from the high line increase character value (all $H_i > 0$) so that the mean of the largest phenotypic class corresponds to $H_1 + H_2 + H_3$, the mean of second largest class to $H_1 + H_2$, etc. This ordering allows us to uniquely estimate the H_i and f_i.



If some of the H_i differ in sign, this ordering breaks down, as illustrated in following example offered by McMillan and Robertson (1974). Setting the mean of the low line at 0, suppose that 23% of the recombinants have value 0, 7% have value 0.5, 41% have value 1.3, and the remaining 29% have value 3.7. Assuming that the effects of all *Q* alleles are positive, the second smallest mean corresponds to $H_1 = 0.5$. Likewise, the third smallest mean corresponds to $H_1 + H_2 = 1.3$, and hence $H_2 = 1.3 - 0.5 = 0.8$. Continuing in this fashion gives the estimated values shown below on the left. In actuality, McMillan and Robertson generated these data using a very different configuration than suggested by the estimate, as shown on the right.



Genetics of *Drosophila* Bristle Number

Drosophila possess a large number of sensory bristles regularly distributed over the adult integument. As discussed in Chapter 15, a number of bristle systems, most notably sterno-

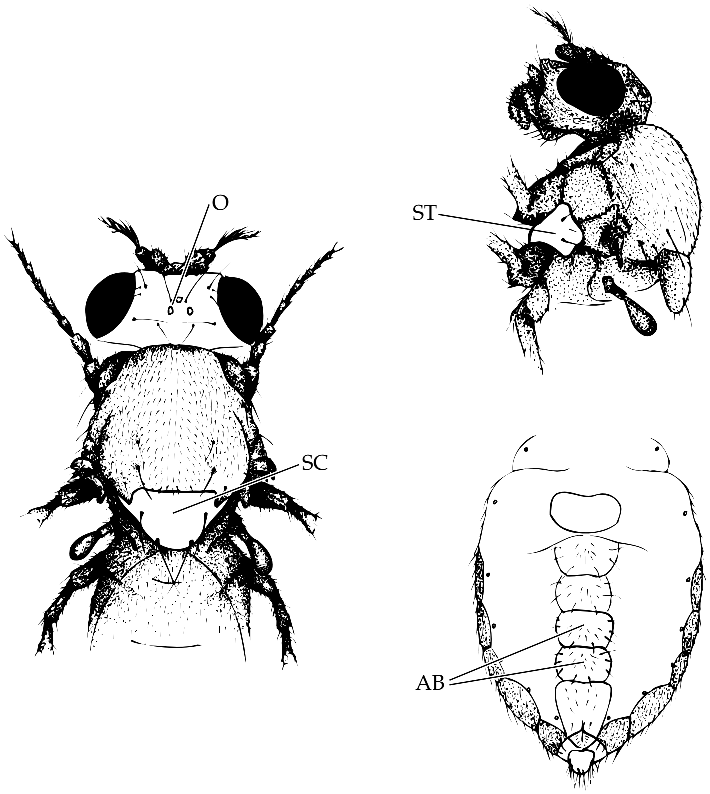


Figure 17.1 Location of head and thoracic bristles in adult *Drosophila*. The most frequently studied bristle systems in quantitative-genetic experiments are sternopleural bristles (ST) on the side of the thorax between the first and second legs, abdominal bristles on the underside of the abdomen (AB), and scutellar bristles (SC) on the upper side of the last thoracic region. Ocellar bristles (O) on the top of the head are also occasionally used.

pleural, abdominal, and scutellar bristles, have been widely used in quantitative-genetic studies (Figure 17.1).

The genetics of sternopleural bristle number have been intensively studied in selected lines using Thoday's method (reviewed in Spickett 1963; Thoday et al. 1964; Spickett and Thoday 1966; Davies and Workman 1971; Davies 1971; Thoday and Thompson 1976; Thoday 1979; Shrimpton and Robertson 1988a, 1988b; Mackay 1995, 1996; Dilda and Mackay 2002; Mackay and Lyman 2004). For example, Spickett and Thoday (1966) examined two selected lines (*vg4* and *vg6*), which showed increases of 15 and 20 bristles, respectively, relative to the standard Oregon R strain. For line *vg6*, 88% of this increase could be accounted for by five factors detectable via Thoday's approach, while three factors (each spanning a distance amounting to less than 2% recombination) accounted for 80% of the increase in line *vg4*. In more detailed analyses restricted to the third chromosome, Davies (1971) and Shrimpton and Robertson (1988a, 1988b) isolated a minimum of 8 and 17 factors, respectively. Extrapolating the latter figure to the rest of the genome yields roughly 60 loci influencing among-line differences in sternopleural bristle number. The distribution of effects for the bristle factors isolated by Shrimpton and Robertson (Figure 17.2) is roughly consistent with results from P-element tagging (Chapter 15), suggesting that sternopleural bristle number is controlled by a few major genes supplemented by numerous genes of smaller effect. Evidence for epistasis was found in most studies.

Spickett (1963) examined the developmental aspects of the isolated factors from line *vg4* in further detail. One factor increased bristle number over the entire fly by increasing

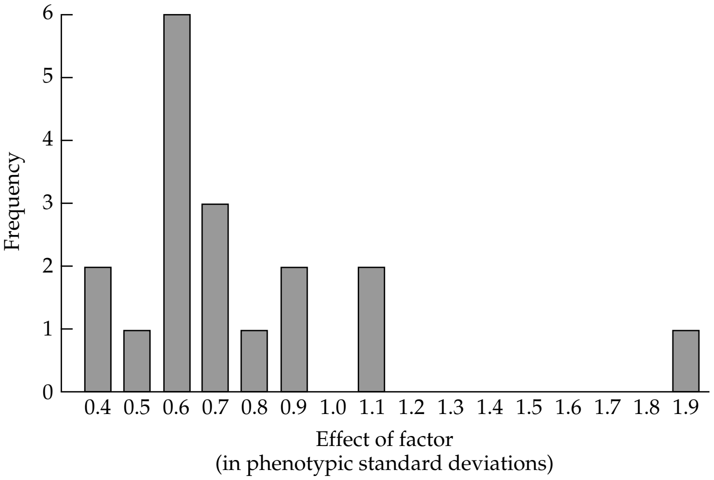


Figure 17.2 Distribution of the effects on sternopleural bristles of *Drosophila melanogaster* third chromosome factors isolated by Shrimpton and Robertson (1988b). This distribution is biased towards large factors, as effects under roughly 0.3 phenotypic standard deviations were not statistically detectable in this study.

total fly cell number. An increase in cell number normally results in a larger fly, but Spickett found a second linked factor that decreased average cell size, resulting in a normal-sized fly with a larger number of cells and more bristles. Two other factors had more local effects. The first was linked to the previous two factors on chromosome 3 and increased the number of bristles in a specific region of the sternopleurite. The second was on chromosome 2 and modified development of a single large bristle into several smaller bristles in this region. Clearly, genetic changes in characters even as superficially simple as bristle number can result from modifications in a variety of developmental pathways.

Genetics of *Drosophila* Speciation

Considerable debate revolves around the genetic basis of speciation. Are such events precipitated by a few genes of major effect (Templeton 1980; Carson and Templeton 1984) or by changes at many loci with individually small effects (Charlesworth et al. 1982; Barton and Charlesworth 1984)? Further, what is the nature of these genetic changes? Are they driven by drift, adaptive selection, or are other forces (such as meiotic drive) involved? Do they involve changes in otherwise relatively normal genes or are unusual genetic features required (e.g., chromosomal rearrangements)? Much of the data on the genetic basis of reproductive isolation due hybrid incompatibly (**postzygotic**, as opposed to pre-mating, or **prezygotic**, barriers) come from crosses between *Drosophila* species that show partial fertility. In discussions of such hybrid incompatibilities, it is important to distinguish between barriers at different stages—**male hybrid sterility** (**HMS**), **female hybrid sterility** (**HFS**), and **hybrid inviablity** (**HI**), which can also be sex-specific—as these different outcomes are most likely the result of different genes. A fairly general pattern in *Drosophila* is that HMS, at least initially, tends to be far more common than HFS or HI (Coyne and Orr 1998, 2004). Before examining the impressive advances in speciation genetics from *Drosophila* studies, it is worth noting that reproductive isolation is not speciation, and the generality of lessons from *Drosophila* remains an open issue (Mallet 2006).

Low-resolution mapping from species crosses provides insight into the number and rough chromosomal locations of hybrid incompatibility factors, while more detailed fine-scale mapping offers some insight into potential underlying genes. The observation that started this line of inquiry was that *D. pseudoobscura* × *D. persimilis* hybrids produce normal females but sterile males. Dobzhansky (1936) backcrossed fertile F_1 hybrid females to

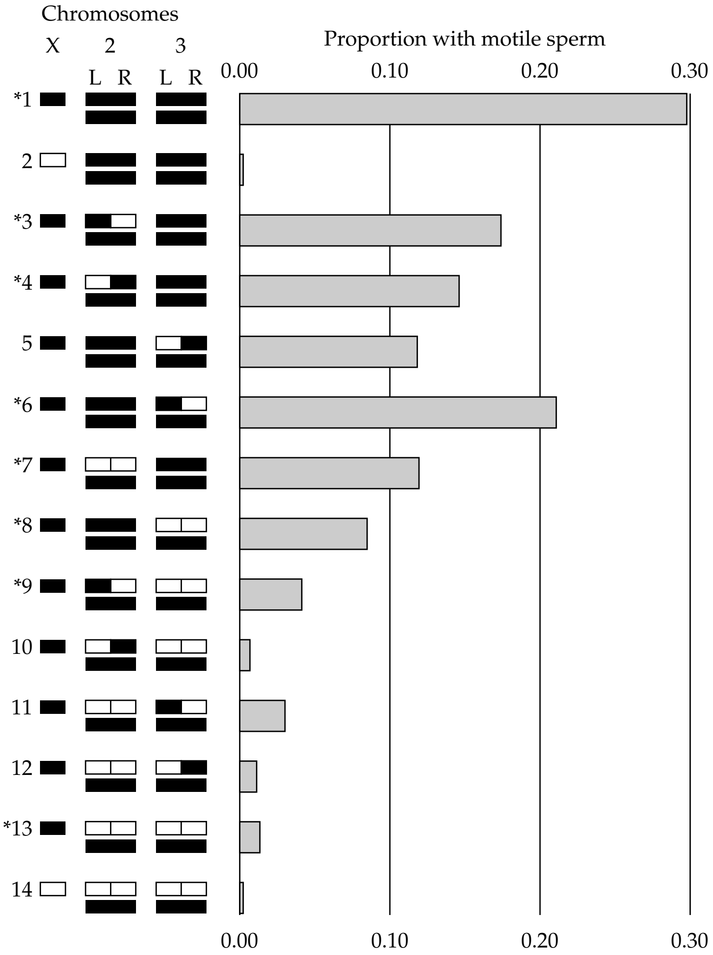


Figure 17.3 Locating male-sterility genes in *D. simulans* × *D. mauritiana* hybrids with chromosomal assays. Fertility is measured by the proportion of males with motile sperm, with chromosome segments from *D. simulans* being shown in black, and those from *D. mauritiana* in white. L and R refer to the left and right arms of chromosomes 2 and 3. All genotypes have a *D. simulans* Y chromosome. Asterisks denote genotypes that produced progeny when crossed to *D. simulans* females. (From Coyne 1984.)

marked males of both species, substituting single chromosomes from one species into an otherwise normal genetic background of the other. He found that testes size (correlated with male fertility) was affected by factors on all five chromosomes, with the strongest effect being on the X chromosome. Orr (1987) found that the reduction in sperm motility in hybrids between these species is largely due to incompatibilities (i.e., epistasis) between factors on the X and Y chromosomes from different species. A finer analysis by Wu and Beckenbach (1983) suggested at least nine factors affecting hybrid male fertility (five X-linked, three autosomal, and one Y-linked).

These studies motivated crosses involving other sibling pairs of *Drosophila* species. For example, Coyne (1984) found that at least five factors are responsible for HMS in *D. simulans* × *D. mauritiana* hybrids, with the main effect again being due to incompatibilities between the X and Y chromosomes of the different species (Figure 17.3). When small sections of the *D. simulans* X chromosome were introgressed into a *D. mauritiana* background by repeated backcrossing, Palopoli and Wu (1994) found four factors influencing HMS, two of which displayed very strong epistasis. Given that they examined less than 20% of the X chromosome, this extrapolates to at least 40 X-linked factors (assuming an equal number of factors were fixed by both species) involved in reproductive isolation between these two species (Davis and Wu 1996). That the genetic units from the above studies are best viewed as *factors*, rather than *single major genes*, is shown by further work from Wu's lab. Introgression of small segments of the X from *D. mauritiana* and *D. sechellia* into *D. simulans* initially suggested that a 500 kilobase region harbored a single major HMS gene (Perez et al. 1993), but a more refined analysis suggested at least two epistatically interacting factors (Perez and Wu 1995). Reed et al. (2008) found that HMS in a cross of *D. arizonae* and *D. mojavensis* was due to multiple epistasic factors mapping to all major chromosomes.

Higher resolution studies were performed by Masly and Presgraves (2007), who introgressed numerous small *D. mauritiana* segments into an otherwise *D. sechellia* background. They made three major conclusions: (i) recessive incompatibilities outnumbered dominant ones; (ii) HMS factors outnumbered all of other types of incompatibilities, and (iii) most X-segment introgressions caused HMS (60%), whereas most autosomal ones did not (18%). Similar observations with other *Drosophila* crosses were made by True et al. (1996) and Tao et al. (2003).

Fine mapping using very small introgressed segments allowed for the localization of several putative **speciation genes**, major factors from one genetic background that tend to negatively interact with numerous other factors from a different background. The classic example is the ***Odysseus*** (***OdsH***) gene, which causes HMS when a 3kb *OdsH* segment (containing exons 3 and 4) from *D. mauritiana* is introgressed into an otherwise completely *D. simulans* background (Sun et al. 2004). This introgression results in the misexpression of several hundred genes, with autosomal genes being more impacted than X-linked ones (Lu et al. 2012). *OdsH* is a homeobox gene (a transcription regulatory factor), which evolved from an ancestral embryonic function into a primarily spermatogenetic one in *D. mauritiana*. Further, *OdsH* shows rapid adaptive evolution in the *D. mauritiana* lineage, with a great excess of replacement over silent substitutions. Such rapid adaptive evolution involving fairly typical genes and a gain (as opposed to loss) of function are common features of most of the handful of isolated genes with large effects on hybrid incompatibility (Orr and Irving 2000; Orr and Presgraves 2000; Orr et al. 2004; Orr 2005).

An important caveat for all of these studies is that they measure the *current* number of factors contributing to reproductive isolation, which can be much greater than the *actual* number involved in the *original* isolation event. Some indications of the potential magnitude of this bias can be seen by considering the classic **Dobzhansky-Muller model** (also refered to as the **Bateman-Dobzhansky-Muller model**) of speciation, which assumes that speciation occurs through epistasis, with isolated populations fixing mutually incompatible alleles (Dobzhansky 1936; Muller 1939; Orr 1996; Turelli and Orr 2000; Turelli et al. 2001; Cutter 2012). This model assumes that a base population fixed for *aabb* is split into two isolated populations, with separate mutations, *A* and *B*, arising in each population. In the *bb* background, *A* has no serious negative fitness consequences, so fixation can occur by drift, giving an *AAbb* population. Likewise, the other population fixes allele *B* to yield *aaBB*, where again *B* has no serious fitness effects in the *aa* background. However, *A* and *B*, when together, are incompatible, such that an *AaBb* hybrid has reduced fitness relative to the ancestral population. A number of such incompatibilities are known in species hybrids (see Orr 1995 for a review), and they are related (at least in principle) to synthetic lethals (Chapter 12). Orr (1995) found that under the Dobzhansky-Muller model, the number of fixed incompatible loci separating the two populations should increase with (at least) the square of separation time, and that incompatibilities involving three or more loci evolve more readily than interactions involving pairs of loci (also see Cabot et al. 1994). This accelerated pace of fixing incompatible loci (a phenomenon known as the **snowball effect**) occurs because the fixation of an incompatibility loci facilitates the fixation of additional such loci. Hence, complex interactions among "speciation genes" are expected to be the norm, rather than the exception. This introduces further complications into the already difficult problem of determining how many factors were actually involved in generating the *initial* isolation event that allowed for the accumulation of further incompatibility genes.

Despite these interpretative difficulties, some general conclusions emerge from these studies. First, in *Drosophila* there is a **large X effect** (also know as **Coyne's rule**; Turelli and Moyle 2007): ***factors with the largest effects on reproductive isolation are located on the X chromosome*** (Coyne and Orr 1989b; Coyne 1992). Although originally framed to include all aspects of reproductive isolation, the large X effect seems to be more general for HMS than for hybrid inviability. The X effect is restricted to reproductive isolation, as no such X chromosome localization is seen for genes influencing morphological differences between *Drosophila* species (e.g., Coyne 1983, 1985; Liu et al. 1996). Consistent with this, Ford and Aquadro (1996) report evidence for selection on X-linked sites, but not autosomal sites, among three semispecies of the *Drosophila athabasca* complex showing reproductive isolation but no obvious morphological differentiation. While an observed excess of HMS factors are seem on the X relative to the autosomes, essential genes for spermatogenesis are actually *underrepresented* on the X relative to autosomes (Masly and Presgraves 2007; Lu et al. 2010). Hence, the large X effect for HMS is not a simply due to the X containing more male fertility factors than the autosomes. One explanation is dominance (see the discussion below on Haldane's rule). Other explanations are related to the divergence in X chromosome dosage compensation in hybrids (Filatov 2018).

The second (more general) conclusion, first noted by Haldane (1922), is that ***when only one sex in a cross-species hybrid is sterile or inviable, it is usually the heterogametic sex*** (e.g., XY males in most animals, WZ females in birds and butterflies). The *Drosophila* data show that male sterility/inviability arises first (often rather quickly), with female sterility/inviablity often developing only after a considerable amount of additional time (Coyne and Orr 1989a, 1997). Much has been written about **Haldane's rule** (**HR**) and its possible causes (e.g., Charlesworth et al. 1987; Orr and Coyne 1989b; Coyne et al. 1991; Frank 1991a; Read and Nee 1991; Coyne 1992; Coyne and Orr 1993; Orr 1993a, 1993b; Virdee 1993; Wu and Davis 1993; Turelli and Orr 1995, 2000; Orr and Turelli 1996; Sawamura 1996; True et al. 1996; Zeng 1996; Laurie 1997; Orr 1997; Turelli and Moyle 2007; Koevoets and Beukeboom 2009; Brothers and Delph 2010; Schilthuizen et al. 2011; Demuth et al. 2014; Delph and Demuth 2016; Moran et al. 2017). While most studies of Haldane's rule examine differences between populations, examining within-population variation is also critical (e.g., Wade et al. 1994; Demuth and Wade 2007; Lachance and True 2010; Cutter 2012). The large-*X* effect and Haldane's rule are often jointly called the **two rules of speciation** (Coyne and Orr 1989b).

While a variety of mechanisms have been proposed to account for HR, Turelli and Orr (1995, 2000; Orr and Turelli 1996) suggested that dominance alone may provide an explanation. If alleles decreasing hybrid fitness are partly recessive, the cumulative effects of sex-linked genes are greater when hemizygous than when heterozygous, resulting in a more pronounced effect in heterogametic hybrids. This effect can be seen by noting that although XX hybrids carry approximately twice the number of deleterious X-linked genes as XY hybrids, if the expression of deleterious effects is less than 50% in heterozygotes, then the overall deleterious effects are smaller in XX hybrids.

While this dominance view for HR appears to be a reasonable explanation for *hybrid inviablity*, a more nuanced view is emerging for *sterility*. It has been suggested that reproductive evolution occurs at a faster rate in males than in females, and that such **faster-male evolution** can either account for, or augment, HMS in species with male heterogametic sex chromosomes (Wu and Davis 1993; Wu et al. 1996; Kulathinal and Singh 2008). Several studies have examined these competing hypotheses. Presgraves and Orr (1998) contrasted HR in the related mosquito genera *Aedes* and *Anopheles*. While *Anopheles* species have traditional heterogametic males (with a largely inactive Y chromosome), *Aedes* species have a single sex-determining locus, but otherwise their Y chromosomes appears to contain the same genes as the X, removing dominance as a potential source for HR. While both HMS and hybrid inviability follow HR in *Anopheles*, Haldane's rule in *Aedes* holds for sterility (HMS occurs, but not HFS), but not for inviablity. In contrast to *Aedes* where both sexes have two active sex chromosomes, in marsupials both sexes are hemizygous, as the paternal X chromosome is uniformly silenced. Watson and Demuth (2012) noted that sterility in marsupials generally follows Haldane's rule (with HMS become more common the HFS), consistent with the faster-male hypothesis given that dominance does not apply. Conversely, Haldane's rule does not seem to apply to inviability. While these observations provide support for the faster-male hypothesis for sterility and the dominance hypothesis for inviablity, other studies suggest the opposite conclusion. Butterflies have females (ZW) as the heterogametic sex and males (ZZ) as the homogametic sex. Crosses between *Heliconius melpomene* and *H. cydno* showed sterile females and normal males, with a large "X" (Z) effect, evidence for the dominance as opposed to the faster-male hypothesis for sterility (Naisbit et al. 2002).

While mapping studies have provided some insight and clarity into the genetics of speciation, they also raise new questions. In particular, what is the role of **meiotic drive** (**segregation distortion**) factors in speciation? Such factors result in a preferential transmission of an allele from heterozygotes, driving them to fixation in the absence of selection (Sandler and Novitski 1957; Jaenike 2001, 2008; Lindholm et al. 2016). Population-genetic models initially suggested a potentially important role for meiotic drive as a speciation process (Frank 1991b; Hurst and Pomiankowski 1991). The idea is that suppresses of meiotic drive factors evolve separately in isolated populations to counter local drive elements. If these suppressers are even slightly recessive, drive is uncovered in hybrids. This hypothesis initially failed to gain traction, as early *Drosophila* crosses did not seem to show the expected drive in hybrids. More recently, however, such segregation distortion has been seen in some hybrids, leading to a reappraisal of its importance (Orr et al. 2006; McDermott and Noor 2010). Tao et al. (2001) localized a factor (which they called *tmy*, for *too much yin*) on chromosome 3 in *D. mauritiana* that impacts both HMS and hybrid male drive in a *D. simulans* background. Similarly, Phadnis and Orr (2009) found a single factor, *Overdrive*, that influences both HMS and segregation distortion in hybrids between the subspecies *D. pseudoobscura pseudoobscura* and *D. p. bogotana*.

GENETIC MAPS

Genetic maps show the ordering of loci along a chromosome and the relative distances between them. As detailed throughout the remainder of this chapter, such maps are essential to the localization of QTLs. The lack of genetic markers historically prevented the construction of detailed maps in all but a few well studied species with short generation times. However, as predicted by Botstein et al. (1980), molecular markers sparked an explosion of genetic maps in humans, economically important plants and animals, and now for just about any species of interest. The theory for constructing genetic maps is highly refined, and we only introduce a few important issues here. For more detailed reviews, the reader should consult Bailey (1961), Lalouel (1992), and the excellent text by Ott (1991). We note the important distinction between a genetic versus a **physical map**. The latter is based on the actual DNA sequence, and hence distance are in base pairs (typically expressed in units of **kilobases**, **kb**; **megabases**, **mb**; or **gigabases**, **gb**; corresponding to 10^3, 10^6, and 10^9 nucleotides, respectively). In contrast, distances on genetic maps measure the *amount of recombination*. While the gene *order* is the same under either map, the *scaling* between loci is not, as rate of genetic recombination is not uniform over a genome.

Map Distances vs. Recombination Frequencies

Genetic map construction involves both the ordering of loci and the measurement of the recombination-based distance between them. Ideally, distances should be additive so that when new loci are added to the map, previously obtained distances do not need to be radically adjusted. Unfortunately, ***recombination* frequencies *are not additive and hence are inappropriate as distance measures***. To illustrate, suppose that three loci are arranged in the order *A*, *B*, and *C* with recombination frequencies c_{AB}, c_{AC}, and c_{BC}, respectively. Each recombination frequency, c, is the probability that an odd number of crossovers occurs between the markers, while $1 - c$ is the probability of an even number (including zero). There are two different ways to get an odd number of crossovers over the interval *A*–*C*: (i) an odd number in *A*–*B* and an even number in *B*–*C*, or (ii) an even number in *A*–*B* and an odd number in *B*–*C*. If there is no **interference**, so that the presence of a crossover in one region has no effect on the frequency of crossovers in adjacent regions, these probabilities can be related as

$$c_{AC} = c_{AB}\,(1 - c_{BC}) + (1 - c_{AB})\,c_{BC} = c_{AB} + c_{BC} - 2c_{AB}\,c_{BC} \tag{17.1}$$

This is **Trow's formula** (Trow 1913). More generally, if the presence of a crossover in one region depresses the probability of a crossover in an adjacent region, then

$$c_{AC} = c_{AB} + c_{BC} - 2(1 - \delta)c_{AB}\,c_{BC} \tag{17.2}$$

where the **interference parameter *δ*** ranges from zero if crossovers are independent (no interference, recovering Trow's formula) to one if the presence of a crossover in one region completely suppresses crossovers in adjacent regions (**complete interference**). Note that recombination frequencies are additive when interference is complete ($\delta = 1$).

Thus, in the absence of very strong interference, recombination frequencies can only be considered to be additive if they are small enough that the product $2c_{AB}c_{BC}$ can be ignored. This is not surprising given that the recombination frequency measures only a part of all recombinant events (those that result in an odd number of crossovers). A genetic map distance m, on the other hand, attempts to measure the ***total number of crossovers*** (both odd and even) between two markers. This is a naturally additive measure, as the number of crossovers between *A* and *C* equals the number of crossovers between *A* and *B* plus the number of crossovers between *B* and *C*.

A number of **mapping functions** attempt to predict the number of crossovers (m) from the observed recombination frequency (c), and, conversely, map a value of m into a value of c. The simplest, derived by Haldane (1919), assumes that crossovers occur randomly and independently over the entire chromosome, i.e., no interference. Let $p(m, k)$ be the probability of k crossovers between two loci that are m map units apart. Under the assumptions of this model, Haldane showed that $p(m, k)$ follows a Poisson distribution (Equation 2.21a), so that the observed fraction of gametes containing an odd number of crossovers is

$$c = \sum_{k=0}^{\infty} p(m, 2k+1) = e^{-m} \sum_{k=0}^{\infty} \frac{m^{2k+1}}{(2k+1)!} = \frac{1 - e^{-2m}}{2} \tag{17.3}$$

where m is the expected number of crossovers. Rearranging, we obtain **Haldane's mapping function**, which yields the (Haldane) map distance m as a function of the observed recombination frequency c,

$$m = -\frac{\ln(1 - 2c)}{2} \tag{17.4}$$

For small c, $m \simeq c$, while for large m, c approaches 1/2. Map distance is usually reported in units of **Morgans** (after T. H. Morgan, who first postulated a chromosomal basis for the existence of linkage groups) or as **centiMorgans** (cM), where 100 cM = 1 Morgan. For example, a Haldane map distance of 20 cM ($m = 0.2$) corresponds to a recombination frequency of $c = (1 - e^{-0.2})/2 \simeq 0.16$. Conversely, an observed recombination frequency of $c = 0.3$ corresponds to an expected number of $m = -\ln(1 - 2 \cdot 0.3)/2 = 0.46$ crossovers (under Haldane's assumptions no interference). Applying Equation 2.21a with $\lambda = 0.49$ expected crossovers gives the probability as 0, 1, 2, or 3 crossovers occurring in this region as 0.631, 0.290, 0.039, and 0.006, respectively.

Although Haldane's mapping function is widely used, several other functions allow for the possibility of crossover interference in adjacent sites (Bailey 1961; Felsenstein 1979;

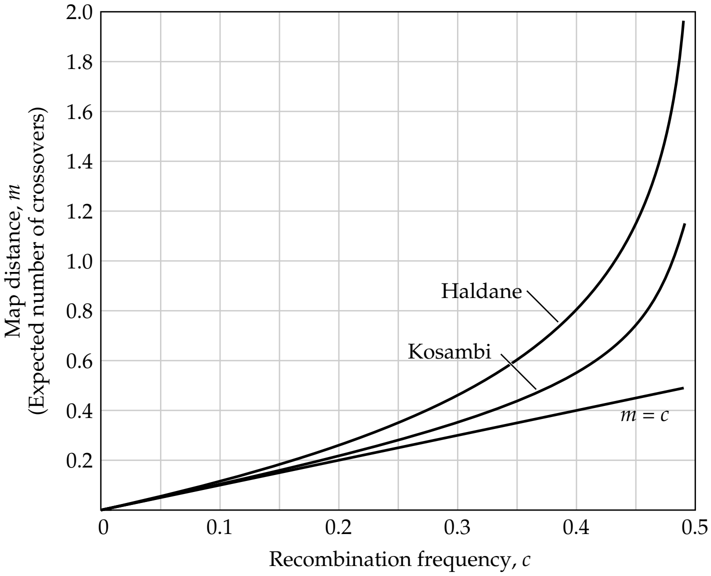


Figure 17.4 Comparison of the Haldane, Kosambi, and simple recombination-frequency ($m = c$) mapping functions, which translate an observed recombination frequency c into an estimate of the expected number of crossovers (the **map distance**, m). The different functions make different assumptions about interference (δ in Equation 17.2). Haldane's function assumes none ($\delta = 0$), Kosambi's assumes a moderate amount, and the simple recombination-frequency model assumes complete interference ($\delta = 1$). For a given c, the Haldane map distance is largest, and the recombination-frequency distance is the smallest. For $c < 0.15$, all three are extremely similar, while the Kosambi and simple function are close for $c < 0.25$.

Karlin 1982; Liberman and Karlin 1984; Pascoe and Morton 1987; Ott 1991; Evans et al. 1993; Foss et al. 1993; Goldstein et al. 1995; Zhao and Speed 1996). For example, geneticists often use **Kosambi's** (1944) **mapping function**, which allows for modest interference,

$$m = \frac{1}{4} \ln \left(\frac{1+2c}{1-2c} \right) \tag{17.5}$$

Damodar Dharmananda Kosambi was an Indian mathematician and polymath, and this 1944 paper was his only one on genetics. Figure 17.4 compares the Haldane and Kosambi mapping functions. For small c (< 0.15), both give $m \simeq c$. For large m, both approach $c = 1/2$. The appropriateness of a potential mapping function can be assessed by checking if the computed map distances depart from additivity (e.g., Pascoe and Morton 1987).

There is no universal relationship between map distance and the actual physical distance between loci (Table 17.2). A centiMorgan can correspond to a span of between ten thousand to millions of nucleotide base pairs, depending on the species. Further, even within a genome, there can be rather dramatic differences. For example, crossovers are often suppressed in particular chromosomal regions (increasing the number of base pairs per cM), such as near the centromere and telomeres (chromosome ends), e.g., True et al. (1996). Further, the rate of recombination is under genetic control with some modifier genes having a general influence throughout the genome and others having a fine-scaled influence on specific chromosomal regions (Brooks 1988). Thus, there may be considerable variation among individuals and among populations in the strength of linkage. The recombination rate can also vary between the sexes. In mammals, for example, females usually have greater map distances than do males. In male *Drosophila*, crossing-over is generally entirely suppressed.

How Many Markers Are Needed?

For many organisms, genetic maps are not yet available, and one must start with randomly

Table 17.2 Basic features of the physical and genetic maps of various eukaryotic groups, derived from a large survey of mapping studies involving high-density molecular markers. The grouping "Other unicellular species" includes algae, apicomplexans, ciliates, kinetoplastids, and oomycetes. Numbers in parentheses denote standard errors, and n denotes the number of species surveyed. Map lengths and mean chromosome (Chr.) sizes are in units of Morgans (M). The last column, kilobases per centiMorgan, is simply the average genome size for the group divided by the average genetic map distance and hence is a fairly crude summary statistic. (From Walsh and Lynch 2018).

Group	Total Map Length	Genome Size (Mb)	Haploid Chr. No.	Mean Chr. Size (M)	Kilobases per centiMorgan
Fungi	18.3 (2.2)	36.4 (3.2)	11.9 (1.2)	1.86 (0.36)	20
Other unicellular sps.	10.9 (1.2)	80.9 (23.3)	12.9 (1.2)	0.96 (0.18)	74
Arthropods	18.1 (3.7)	679.6 (172.4)	16.1 (3.4)	1.20 (0.18)	375
Mollusks	9.2 (1.1)	1270.7 (177.2)	13.3 (1.6)	0.71 (0.09)	1313
Nematodes	4.5 (1.2)	97.6 (2.5)	7.3 (1.3)	0.59 (0.05)	217
Fish	16.0 (2.3)	1185.4 (190.5)	25.1 (0.6)	0.63 (0.08)	741
Birds	23.1 (5.4)	1334.0 (48.6)	39.6 (0.4)	0.58 (0.14)	577
Mammals	23.9 (2.5)	3222.0 (108.1)	22.1 (2.2)	1.10 (0.07)	1348
Angiosperms	15.9 (1.6)	2020.3 (434.2)	13.2 (0.9)	1.19 (0.07)	1270

obtained markers. As a rough approximation for the number of random polymorphic markers necessary for a desired saturation of the linkage map, assume a circular linkage map of total length L map units. In order to have a fraction p of all loci within m map units of some marker, the required number of randomly distributed markers is

$$n = \frac{\ln(1-p)}{\ln(1-2m/L)} \tag{17.6a}$$

(Lange and Boehnke 1982; Beckmann and Soller 1983). This rearranges to give the proportion of the genome within m map units of a marker as approximately

$$p \simeq 1 - \exp\left(-\frac{2mn}{L}\right) \tag{17.6b}$$

(Jacob et al. 1991). Assuming a circular genome map ignores the effect of chromosome ends, causing Equation 17.6a to underestimate n and Equation 17.6b to overestimate p. A more general expression, which accounts for linear chromosomes, was given by Bishop et al. (1983). Assuming a haploid chromosome number C,

$$p = 1 - \frac{2C\left[(1-x)^{n+1} - (1-2x)^{n+1}\right]}{n+1} - (1-2x\,C)\,(1-2x)^n \tag{17.6c}$$

where $x = m/L$.

An alternative consideration is the average distance between a locus and the nearest random marker. Martin et al. (1991) showed that the expected distance of a gene from the closest of n random markers is

$$E[\,m\,] = \frac{L}{2(n+1)} \tag{17.7a}$$

with the upper 95% confidence interval for this distance given by

$$\frac{L}{2}\left(1 - 0.05^{1/n}\right) \tag{17.7b}$$

It is important to note that the above expressions refer to *randomly chosen* markers. It is much more efficient to use sets of equally spaced markers. If the entire genome is covered by marker loci equally spaced at m map units apart, no locus is more than $m/2$ from any marker, and the average distance of a locus from a marker (assuming a uniform distribution of loci along the chromosome) is $m/4$. Note that evenly-spaced markers refer to distances measured on a recombinational, rather than physical distance, scale. Hence, even with a complete reference genome (physical maps) in hand, additional information (localized recombination rates) is required to evenly space the markers.

Example 17.3. The human linkage map is approximately 33 Morgans long with haploid chromosome number $C = 23$. How many random markers are required to achieve a 90% probability that at least one marker is within 10 map units (centiMorgans) of a randomly chosen gene?

Here L = 3300 cM, m = 10 cM, and p = 0.9. The circular-chromosome approximation (Equation 17.6a) gives

$$n = \frac{\ln(1-.9)}{\ln(1-2/330)} \simeq 379$$

Numerically solving Equation 17.6c with $p = 0.9$ gives n = 404. Hence, the effect of ignoring chromosome ends underestimates n by about 7%. Suppose that 110 random markers are used. From Equation 17.7a, the expected distance of a particular gene from the nearest of these 110 markers is 16.9 cM, while Equation 17.7b gives the upper 95% confidence interval for this distance as 44.3 cM. In contrast, suppose these 110 markers are not random, but are instead chosen to be equally spaced at 30 cM apart. With this spacing, no locus is more than 15 cM from a marker, and loci are, on average, 7.5 cM from a marker.

DESIGN STRATEGIES FOR DETECTING MARKER-TRAIT ASSOCIATIONS

We now turn our attention to the mapping of individual QTLs (the causative loci that underlie variation in a focal trait). The distinction is often made been **linkage mapping**, exploiting the linkage disequilibrium (LD) generated by an excess of parental gametes *among relatives* (Example 5.6), and **LD mapping** using *a population sample of nonrelatives*. The former allows for QTL mapping even when markers are only loosely linked to causative loci (and hence only a few dozen to a few hundred markers are needed). The latter requires very dense markers (thousands to millions), as population-level LD generally only occurs between very tightly-linked loci (Chapter 5). Given the limitations on the number of available markers, early QTL mapping attempts were linkage-based. More recently, the vast abundance of makers has led to a shift toward population-level QTL mapping. A variety of experimental designs using linkage information in either a cross or a pedigree have been proposed, and we consider these in some detail in Chapters 18 (inbred line crosses) and 19 (outbred crosses and pedigrees), while designs using population-level linkage disequilibrium are examined in Chapter 20. Since those chapters focus on methods for estimating both QTL effects and map position, we defer further discussion of most statistical issues until then. The remainder of this chapter is concerned with general design strategies for mapping.

That QTLs, like normal genes, can be mapped (assigned to linkage groups) was first demonstrated by Payne (1918), who found that the X chromosome from selected lines of *Drosophila* contains multiple factors influencing scutellar bristle number. Over the next two decades, a number of workers found associations between Mendelian markers (i.e., distinct phenotypes associated with known single-locus genotypes) and quantitative traits in crosses between inbred lines. For example, Sax (1923) crossed two inbred bean lines (*Phaseolus vulgaris*) differing in seed pigment and weight, with the pigmented parents having heavier seeds than the nonpigmented parents. These crosses demonstrated that seed pigment is determined by a single locus with two alleles, *P* and *p*. Among F_2 segregants from this cross, *PP* and *Pp* seeds were, respectively, 4.3 ± 0.8 and 1.9 ± 0.6 centigrams heavier than *pp* seeds. Hence the *P* allele was linked to a factor (or factors) that act in an additive fashion on seed weight. In a similar manner, Lindstrom (1924) demonstrated linkage between a locus for fruit color and factors for fruit size in tomatoes, while Smith (1937) observed associations between corolla size differences and several independently segregating flower color genes in tobacco. Similar early studies were reported in maize (Lindstrom 1931), peas (Rasmusson 1927), barley (Wexelesen 1933, 1934), and mice (Green 1931, 1933).

Despite just using single marker loci, these early studies raised the possibility that QTLs could be mapped with some precision, perhaps even allowing for the characterization of individual loci, given a population whose QTLs are in linkage disequilibrium with scored markers. Disequilibrium between marker loci and QTLs is the key requirement, as it creates **marker-trait associations**, with different marker genotypes having different expected values for characters influenced by QTLs linked to, and in LD with, these markers. This simple idea is the foundation for QTL mapping. Creating populations with loci in LD, even when they are only modestly linked, is straightforward: crosses between inbred lines have this property, as do sibs and other collections of relatives (Example 5.6). For mapping purposes, crosses between inbred lines have the fewest complications. Consider the cross of between $MMQQ$ and $mmqq$ inbred lines, where M and m are scored marker alleles linked to a QTL with alleles Q and q. The F_1 progeny (who are all MQ/mq) from such a cross display maximum disequilibrium and, being genetically identical, are equally informative as parents. As will be examined in detail in Chapter 18, a variety of mapping populations (such as F_2s or backcrosses) can be generated by using such F_1 parents,.

A number of problems, which are elaborated on in Chapter 19, conspire to make linkage-based QTL mapping considerably more difficult in outbred-line crosses. Not all parents are guaranteed to be informative (i.e., are marker-QTL double heterozygotes), as some may be segregating marker alleles, but not linked QTLs ($MmQQ$ or $Mmqq$), or vice-versa (e.g., $MMQq$ or $mmQq$). Further, the marker-QTL linkage phase can differ between different sets of relatives. For example, marker allele *M* might be associated with QTL allele *q* in one family, but with *Q* in another. Thus, marker-trait associations have to be assessed in each set of relatives, rather than (as with inbred-line crosses) averaged over all individuals.

QTL mapping based on the LD among tightly-linked loci in a random population sample requires orders of magnitude more markers than linkage-based approaches. However, it also has two significant advantages (Chapter 20). First, mapping based on crosses (linkage mapping) is limited by family size (or pedigree depth), while no such limitation occurs when using population-level LD mapping. Hence, the number of individuals phenotyped is usually on the order of a few hundred in linkage-based studies, but thousands to hundreds-of-thousands in population-level studies. Second, population-level studies have significantly greater mapping resolution, usually on the order of kilobases (i.e., nearly single-gene scale), than linkage-based studies, which typically have only megabase resolution (large chromosomal segment scale). LD block sizes in the former are much smaller as they are shaped by a great many generations of recombination, as opposed to one or two generations of recombination in most designs for the latter (Equations 5.13d and 5.16c).

Example 17.4. Likelihood functions for QTL mapping follow from standard mixture models (Chapter 16). Consider the simple backcross design, wherein two completely inbred lines (with marker/QTL genotypes *MMQQ* and *mmqq*) are crossed to form an F_1 (*MQ/mq*) which is then backcrossed to the *MMQQ* population. If c denotes the marker-QTL recombination frequency, then a fraction $(1-c)$ of *M*-bearing F_1 gametes contain *Q*, while c contain *q*. Likewise $(1-c)$ and c of *m*-bearing gametes contain *q* and *Q*, respectively. Because the gamete from the parental population is always *MQ*, the conditional probabilities of QTL genotypes given marker genotypes are

$$\Pr(\,QQ\,|\,MM\,) = \Pr(\,Qq\,|\,Mm\,) = 1 - c$$

$$\Pr(\,Qq\,|\,MM\,) = \Pr(\,QQ\,|\,Mm\,) = c$$

Hence, if z_j is the character value for individual j, the likelihood depends on the marker genotype,

$$\ell(z_j) = \begin{cases} (1-c)\cdot\varphi(z_j,\mu_{QQ},\sigma^2) + c\cdot\varphi(z_j,\mu_{Qq},\sigma^2), & \text{if marker} = MM \\ \\ c\cdot\varphi(z_j,\mu_{QQ},\sigma^2) + (1-c)\cdot\varphi(z_j,\mu_{Qq},\sigma^2), & \text{if marker} = Mm \end{cases}$$

As in Chapter 16, we have assumed that phenotypes are normally distributed with potentially different means for each genotype but a common variance, with $\varphi(z,\mu,\sigma^2)$ denoting the density function for a normal distribution with mean μ and variance σ^2. The total likelihood for n measured backcross individuals is the product of individual likelihoods, $\prod \ell(z_j)$. The maximum-likelihood estimates of the four model parameters (μ_{QQ}, μ_{Qq}, c, σ^2) are obtained by maximizing the likelihood with respect to these variables, treating the observed z_j as fixed constants.

Hypothesis testing follows by standard likelihood-ratio tests (Chapters 16 and 18; Appendix 4). The appropriate test for the presence of a QTL linked to the marker compares the likelihood under the assumed full model with the likelihood under a model of no QTL, in which each individual character value is assumed to be drawn from the same normal with unknown mean μ and variance σ^2. The maximum of this restricted likelihood is given by Equation 16.8. The resulting likelihood-ratio test has $4-2=2$ degrees of freedom (four free parameters in the full model; μ and σ^2 in the restricted model). Knott and Haley (1992a) showed that this test is not biased by the presence of unlinked QTLs. It is, however, biased if one or more additional QTLs are linked to the marker under consideration.

An observed association between alleles at a polymorphic marker locus and the value of a quantitative trait can result either because of gametic-phase disequilibrium between the marker locus and a QTL or because the marker itself has a pleiotropic effect on the trait. A simple test for pleiotropy versus disequilibrium is to examine a marker-trait association over several generations of random mating. An association due to pleiotropy will not decay over time, while one due to linkage alone will. A caveat is that if linkage is very tight ($c \ll 0.5$), it can take many generations before any substantial decay is observed (Equation 5.16c). If one suspects that the marker itself is the QTL (which, of course, is highly unlikely), a variety of **candidate-locus** approaches (discussed below) can be used to test this hypothesis.

Example 17.5. Following crosses between inbred lines of barley, Powell et al. (1985a, 1985b) found that two loci with recessive dwarfing alleles (*ert* and *denso*) were associated with other quantitative traits: *ert* with reduced seed weight and *denso* with reduced height. By crossing different lines to produce an F$_2$, it was found that the *ert* locus accounted for roughly 84% of the additive genetic variation in seed weight after both one and three generations of recombination, suggesting that the association between *ert* and seed weight is either due to pleiotropy or tight linkage. For example, the largest value of c that retains 90% of the originally association satisfies $(1-c)^3 = 0.90$, or $c = 1 - 0.90^{1/3} \simeq 0.03$. In contrast, the *denso* locus accounted for 58% of the additive genetic variance for height in lines undergoing a single round of recombination but only 35% in lines undergoing three rounds. Hence, for the *denso* locus, much of the initial association was due to gametic-phase disequilibrium.

Selective Genotyping and Progeny Testing

Historically, markers were more expensive to score than phenotypes. A strategy proposed to handle this setting when our interest is in a single trait is to first phenotype a number of individuals and then genotype only a *selected subset* of these. Known as **selective genotyping**, this strategy can result in a large increase in power for the simple reason that much of the linkage information resides in individuals with extreme phenotypes (Lebowitz et al. 1987; Lander and Botstein 1989; Carey and Williamson 1991; Darvasi and Soller 1992). However, while selective genotyping offers increased power to *detect* a QTL, it also produces biased estimates of their effects. Unbiased estimates can be obtained by maximum likelihood, either using a likelihood function that directly accounts for the sampling bias (Darvasi and Soller 1992; Muranty and Goffinet 1997; Xu and Vogl 2000) or by treating unscored genotypes as missing values (Lander and Botstein 1989; Johnson et al. 1999). The irony is that marker technology has now advanced to the point where *genotyping is often faster and cheaper than phenotyping*. Indeed, in many settings the bottleneck in marker-trait association studies is the gathering of highly accurate measurements on phenotypes.

This raises the important point that ***marker-trait associations are only as good as the quality of phenotyping***. Usually the reliability of marker-genotype scoring can be assessed with a number of quality control checks (such as departures from Hardy-Weinberg proportions). However, such checks are usually not available for phenotypes. Poor-quality phenotyping results, at a minimum, in a reduction of power. Hence, schemes that improve the precision of phenotyping (such as using good experimental design practices; Appendix 9) improve the power and precision of QTL mapping.

In those organisms where asexual propagation is possible, **replicated progeny** can be used to improve phenotyping precision (Cowen 1988; Simpson 1989; Lander and Botstein 1989; Soller and Beckmann 1990; Knapp and Bridges 1990). The idea here is to reduce the effects of environmental variance by asexually replicating each genotype, using the mean values of these replicated progeny in place of individual values. This is an efficient strategy in that if the heritability of the character is low, even scoring only a few replicated progeny can result in a significant increase in power. Soller and Beckmann (1990) showed that most of the increase in power occurs by measuring 10 or fewer replicated progeny. Besides offering an increase in power, progeny replication allows marker-trait associations to be examined across environments, providing a basis for estimating QTL × environment interactions (Chapter 27). A variant approach is use $F_{2:3}$ (or related) families in place of replicated progeny. Here, one genotypes an F_2 individual, which is then selfed, with the mean of that individual's selfed progeny used as their trait value.

Recombinant Inbred (RILs) and Doubled Haploid (DHLs) Lines

When asexual propagation is not possible, a related method improve the precision of phenotyping is to use **recombinant inbred lines** (**RILs**). These can be constructed for any organism by taking an F_1 line through multiple rounds of selfing (e.g., Burr and Burr 1991) or multiple generations of brother-sister mating (e.g., Bailey 1981). The resulting lines have essentially no within-line genetic variance (ignoring new mutation and small amounts of residual heterozygosity), whereas the genetic variance among lines is considerable, as each RIL represents a different multilocus genotype. This was the approach used by Mackay and colleagues to study the difficult phenotype of lifespan in *Drosophila* (e.g., Nuzhdin et al. 1997; Ivanove et al. 2015). Longevity is a trait with considerable noise when based on a single measurement (that for a single individual), while the average life span for a RIL of a given genotype can be measured with considerable precision.

A related approach to RILs is the use of **doubled-haploid lines** (**DL** lines or **DHLs**). These are constructed using a variety of reproductive tricks that take a cell with a single active haploid gamete and then doubles its chromosome number, instantly producing completely homozygous individuals. While DH technology was originally restricted to just a few plant systems (in particular, corn), it has now been widely expanded in plants, with variants of this technology now available for several hundred species (Maluszynski et al. 2003; Forster et al. 2007; Touraev et al. 2009). Following the pioneering work of Streisinger et al. (1981) in zebra fish, DH lines have been constructed for a number of commercial fish species (Franěk et al. 2020; Hansen et al. 2020). In fish, most are **gynogenetic DHLs**, produced by **gynogenesis** (inactivation of sperm DNA through irradiation or chemical treatment and doubling the haploid egg fertilized with such sperm). However, there are also

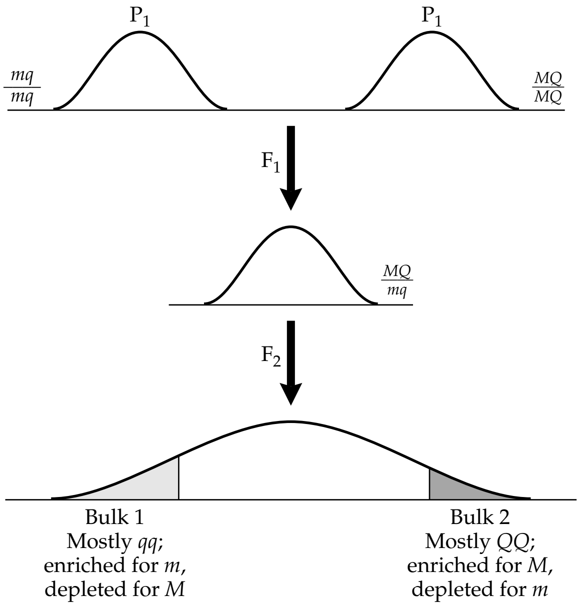


Figure 17.5 Bulked segregant analysis for mapping QTLs. Assume that a major QTL with alleles Q/q is tightly linked to a marker locus with alleles M/m. If we cross two inbred populations (*MMQQ* and *mmqq*) and then sample the extreme tails of the F$_2$ phenotype distribution, the lower tail is enriched for *qq*, and the upper tail is enriched for *QQ*. Since the marker is closely linked to the QTL, the lower tail is enriched for marker allele *m* and depleted for *M*, while the opposite is true in the upper tail.

androgenetic DHLs, produced by **androgenesis** (inactivation of the nuclear DNA in the haploid egg, and doubling of the fertilized egg). Typically, the production of DHLs involves only a single round of recombination, while RILs created by selfing have around two effective rounds of recombination (accounting for the rapid reduction in heterozygotes), while RILs from brother-sister mating have around four effective generations (Equation 18.5b). Hence, DHLs tend to have poorer mapping resolution than RILs.

Once the considerable work to generate and molecularly characterize a set of RILs or DHLs has been done, any character can be examined, and the previous marker information used to look for marker-trait associations. Only one laboratory needs to characterize the marker genotypes for the set of lines, while subsequent investigators (with these genotypes in hand) can look for marker-trait associations across the set of lines by simply phenotyping traits of interest. Like asexual lineages, RILs and DHLs offer a particularly easy approach for measuring QTL × environment interactions (Chapter 27), since the same lines can be raised over different sets of environments. For these reasons, behavioral geneticists typically use RILs for mapping behavioral QTLs in mice (e.g., Plomin et al. 1991; Belknap et al. 1993; Crabbe et al. 1994; Flint 2003).

Bulked Segregant Analysis (BSA)

A variant of selective genotyping is **bulked segregant analysis** or **BSA** (Arnheim et al. 1985; Michelmore et al. 1991; Zou et al. 2016). Here individuals are combined (**bulked** or **pooled**) into groups based on trait value (Figure 17.5), with the marker frequencies in each bulk then estimated and contrasted. Marker alleles in linkage disequilibrium with QTL alleles are expected to have a nonrandom distribution across bulks, in the extreme having a particular marker allele present only in one bulk and the alternative allele present only in the other bulk (Hill 1998b). Unlinked markers (and marker alleles in linkage equilibrium with QTLs) are expected to be randomly distributed across bulks. Figure 17.6 shows an example of this approach, mapping four major genes with a resolution of less than 2 Mb using bulks formed

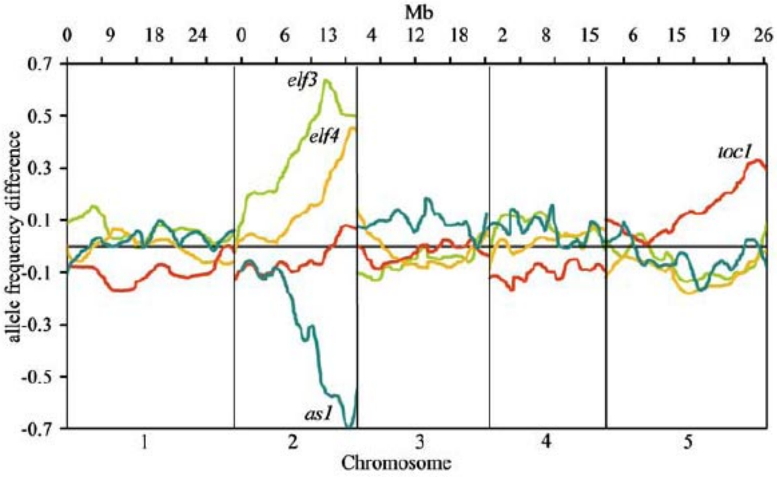


Figure 17.6 Using BSA to map mutations impacting circadian clock and developmental traits in F_2 bulks of *Arabidopsis*. Spline-smoothed curves are plotted of allele-frequency differences between the upper and lower bulks based on hybridization strengths on an Affymetrix gene chip. Four different bulked pairs (represented by the four separate curves) were used to detect mutations in the *early flowering 3* (*efl3*), *early flowering 4* (*efl4*), *time of CAB expression 1* (*toca*), and *asymmetric leaves 1* (*as1*) genes that impacted the bulked trait. (After Hazen et al. 2005.)

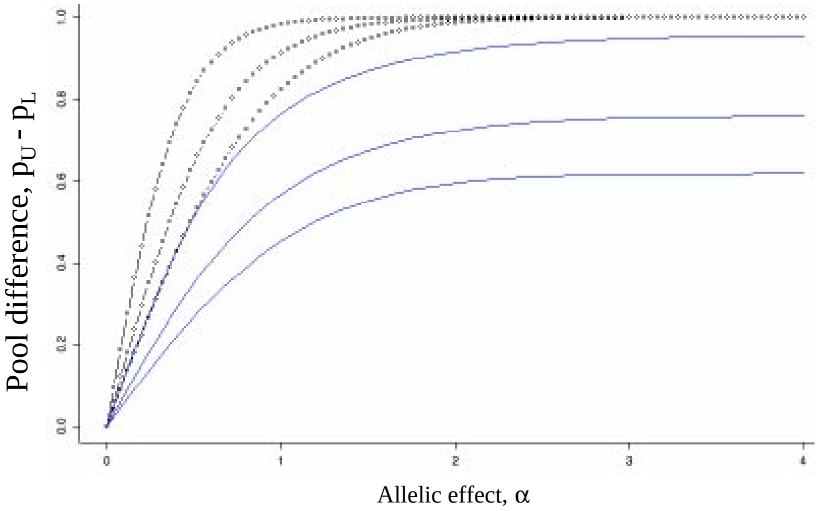


Figure 17.7 The expected allele QTL frequency difference between the upper and lower bulks of a BSA as a function of the effect (α, in standard deviations) of an additive QTL and the bulk thresholds (δ, which loosely corresponds to the number of standard deviations above and below the mean). The upper three (dotted) lines are the values for RILs and (from upper to lower) corresponds to δ values of 2, 1, and 0.5, while the lower three (solid) lines correspond to an F_2-based BSA with the same ordering for these three δ values. Details in Example 17.8.

from F_2 populations in *Arabidopsis*.

Early attempts to quantify allele frequency differences between pools used indirect approaches such as optical density of allelic bands (Pacek et al. 1993; Khatib et al 1994; Darvasi and Soller 1994a). The extension of this approach in the genomics era uses quantification via differential hybridization based on either transcriptomics (Borevitz et al. 2003; Brauer et al 2006; Pandit et al 2010) or SNP chips (Hazen et al. 2005; Wenger et al. 2010; Becker et al. 2011). The application of array scoring (RNA or SNPs) in BSA has been called **eXtreme array**

mapping, **XAM** (Woyln et al. 2004). A more direct approach is direct counting of either the number of allelic sequences or the number of transcripts in each pool via **next generation sequencing**, **NGS**, very high coverage sequencing of the pool (Magwene et al. 2011; Park et al. 2014). NGS data generates a series of counts for each pool, generating a 2×2 contingency table (with rows corresponding to the two alternate marker alleles and columns to the two pools). Magwene et al. (2011) proposed a modification of the G goodness-of-fit statistic (Appendix 4) to test for significance with such NGS data. A power analysis by these authors found that the optimal design was to use rather modest thresholds (around 20%) for bulking phenotypes, as while allele frequencies would be more differentiated using more extreme thresholds, this is countered by fewer members in these pool, compromising power.

Because bulked segregant analysis requires significant enrichment of the alternative QTL alleles in the tails of the trait distribution, it is expected have poor power for detecting QTLs of small to modest effect (Figure 17.7; Example 17.8). It also requires using either derivatives from a single inbred-line cross (such as F_2s or RILs) or progeny from very large segregating families, as the method fails if groups with the marker and QTL alleles in different linkage phases are lumped. In spite of these limitations, BSA is so straightforward, allowing rapid screening of a very large number of markers, that it is reasonable to consider it in many cases. As shown in Figure 17.7, this approach is most powerful when using RILs. This is due to RIL pools only consisting of QTL homozygotes, while heterozyogtes are present when F_2s are bulked. A further advantage of using RILs is that the replication of these lines provides very precise estimates of their genotypic values (Venuprasad et al. 2009; Vikram et al. 2012). One constraint on using BSA is that a bulk is *trait specific*, so that if one is attempting to map QTLs for multiple traits, independent bulks for each trait must be formed.

Finally, we note that BSA can be used to find *additional markers* linked to a particular region (Giovannoni et al. 1991). Sorting individuals from a cross into separate pools based on alternate alleles at a marker enriches those pools for additional linked markers. Thus, bulking on a marker of interest provides for very efficient screening of a large number of markers. Such an increase in marker density is required for finer mapping of QTL position.

Example 17.6. Yaghoobi et al. (1995) used bulked segregant analysis to map a major gene for root-knot nematode resistance in tomatoes. Forty-eight backcross individuals from a cross between resistant and susceptible strains (F_1 back-crossed to the susceptible parent) were scored, 25 of which were susceptible, and 23 resistant. This 1:1 segregation ratio is consistent with the hypothesis that a single dominant gene underlies resistance. To map this gene, two bulks of DNA, based on six resistant and six susceptible backcross plants were formed. Each bulk was screened for RAPDs using 520 different 10-base primers. Each primer gave on average about eight bands, resulting in 4,160 bands being scored in each pool, about 3% of which varied significantly between pools. These significant primers were then used to probe each of the original 48 backcross plants. One marker was present in 20 of 23 resistant plants and none (out of 25) of the susceptible plants, suggesting linkage to a major resistance gene.

Example 17.7. As a test of BSA, Mansur et al. (1993) examined four traits—maturity, plant height, lodging (a measure of plant structure), and seed yield—in 284 recombinant inbred lines generated by a cross between two soybean (*Glycine max*) cultivars. The four chosen traits were measured in each of the 284 RILs grown in two distinct environments (Minnesota and Chile) and the 20 highest- and lowest-performing lines, averaged over both environments, were selected for each trait. DNA was extracted from each of the extreme lines and bulked into a high and low sample for each of the traits. The resulting DNA was tested using radioactive probes for RFLP markers, and the amount of hybridization to each probe was quantified using a phosphoimager. The authors had previously used RFLPs to map a number of QTLs for these and several other traits using maximum-likelihood interval mapping (Chapter 18) in F_2 families. The previous RFLP marker–QTL associations were confirmed, and one marker that showed marginal linkage to maturity and height under interval mapping showed very strong linkage to these traits as well as to lodging and yield.

Example 17.8. What is the expected QTL allele frequency difference between the two pools in a BSA? Consider a trait, z, whose mean has been centered at zero. Assume an additive QTL, whose phenotypic distributions for the trait value, z, are normal, with

$$z\,|\,QQ \sim \mathrm{N}(\alpha,1), \qquad z\,|\,Qq \sim \mathrm{N}(0,1), \qquad z\,|\,qq \sim \mathrm{N}(-\alpha,1)$$

Recalling Equation 15.3, the resulting phenotypic distribution for z is a mixture, with a variance of $1+\sigma^2(\mu_G)$. If (as it often the case), the BSA population is the F_2 between two inbred lines, then freq(Q) = 1/2 and $\sigma^2(\mu_G) = 2(1/4)\alpha^2$, for a trait standard deviation of $\sigma_z = \sqrt{1+\alpha^2/2} \simeq 1$ for $\alpha < 1$. For RILs, $\sigma^2(\mu_G) = 2(1/2)\alpha^2 = \alpha^2$.

Suppose we contrast the frequency of Q between an upper bulk, U, that consists of all measured individuals whose trait values are at least δ above the mean ($z \geq \delta$), and a lower bulk, L, of individuals whose trait values are less that δ below the mean ($z \leq -\delta$). The expected frequency of Q in the upper bulk is

$$p_U = \Pr(QQ\,|\,z \geq \delta) + (1/2)\Pr(Qq\,|\,z \geq \delta)$$

while its expected frequency in the lower bulk is

$$p_L = \Pr(QQ\,|\,z \leq -\delta) + (1/2)\Pr(Qq\,|\,z \leq -\delta)$$

As shown in Example 3.2, one can use Bayes theorem to obtain expressions for these various probabilities for QTL genotypes conditioned on trait value. Consider QQ first. Applying Bayes theorem (Equation 3.3b),

$$\Pr(QQ\,|\,z \geq \delta) = \frac{\Pr(z_{QQ} \geq \delta)\Pr(QQ)}{\Pr(z_{QQ} \geq \delta)\Pr(QQ) + \Pr(z_{Qq} \geq \delta)\Pr(Qq) + \Pr(z_{qq} \geq \delta)\Pr(qq)}$$

Letting $u \sim N(0,1)$ denote a unit normal, note that $z_{QQ} = u+\alpha$, $z_{Qq} = u$, and $z_{qq} = u-\alpha$. The above expression (for an F_2) reduces to

$$\frac{\Pr(\,u+\alpha \geq \delta\,)}{\Pr(\,u+\alpha \geq \delta\,) + 2\Pr(\,u \geq \delta\,) + \Pr(\,u-\alpha \geq \delta\,)}$$

$$= \frac{\Pr(\,u \geq \delta-\alpha\,)}{\Pr(\,u \geq \delta-\alpha\,) + 2\Pr(\,u \geq \delta\,) + \Pr(\,u \geq \delta+\alpha\,)}$$

Hence,

$$p_U = \frac{\Pr(\,u \geq \delta-\alpha\,) + 2(1/2)\Pr(\,u \geq \delta\,)}{\Pr(\,u \geq \delta-\alpha\,) + 2\Pr(\,u \geq \delta\,) + \Pr(\,u \geq \delta+\alpha\,)} \tag{17.8a}$$

Similarly,

$$p_L = \frac{\Pr(\,u \leq -\delta-\alpha\,) + \Pr(\,u \leq -\delta\,)}{\Pr(\,u \leq -\delta-\alpha\,) + 2\Pr(\,u \leq -\delta\,) + \Pr(\,u \leq -\delta+\alpha\,)} \tag{17.8b}$$

For BSA based on RILs from an F_2, the heterozygote terms, $\Pr(\,u \leq -\delta\,)$ and $\Pr(\,u \geq \delta\,)$, are absent, resulting in higher power (Figure 17.7), with

$$p_U - p_L = \frac{\Pr(\,u \geq \delta-\alpha\,)}{\Pr(\,u \geq \delta-\alpha\,) + \Pr(\,u \geq \delta+\alpha\,)} + \frac{\Pr(\,u \leq -\delta-\alpha\,)}{\Pr(\,u \leq -\delta-\alpha\,) + \Pr[\,u \leq -\delta+\alpha\,)} \tag{17.8c}$$

These expressions can be easily computed using the standard normal cumulative probability function found in any statistical package. For example, if we contrast the upper and lower 20% of the population (which for $\sigma_z^2 \simeq 1$ corresponds to δ = 0.84), the between-bulk difference in allele frequencies, $p_U - p_L$, becomes 0.070, 0.323, 0.532, and 0.683 for α = 0.1, 0.5, 1, and 2, respectively for an F_2 bulk, increasing to values of 0.139, 0.606, 0.890, and 0.995 for a BSA using RILs. Figure 17.7 plots this between-bulk difference for various values of α and δ for both F_2 and RIL designs. A final adjustment is that the above calculations assumed a focus on the frequency of the causative allele, Q, while in fact we are usually following a marker allele, M, linked to the QTL at recombination fraction c. The expected allele frequency change for the marker follows with α replaced by $\alpha(1-2\,\widetilde{c}\,)$, where

$$\widetilde{c} = \begin{cases} c & \text{for DHLs, } F_2\text{s} \\ 2c/(1+2c) & \text{for RILs formed by selfing} \\ 4c/(1+6c) & \text{for RILs formed by brother-sister mating} \end{cases} \tag{17.8d}$$

QTL Mapping by Marker Allele-frequency Changes in Populations under Selection

Selective genotyping amounts to a single-generation selection experiment, as individuals are chosen from the tails of the character distribution. As originally suggested by Stuber and Moll (1972) and Stuber et al. (1980), this mapping approach can be generalized by looking for shifts in marker allele frequencies following several generations of selection on a focal trait. Marker loci initially in linkage disequilibrium to a QTL for that trait will increase (decrease) in frequency as linked QTL alleles increase (decrease) due to selection. In the case of Stuber and colleagues, they searched for shifts in allozyme frequencies in lines of maize selected for yield. A slightly more powerful design is to divergently select a base population for several generations and then test for significant changes in marker allele frequencies between the up- and down-selected lines (Nuzhdin and Pasyukova 1991; Keightley and Bulfield 1993; Nuzhdin et al. 1993; Ollivier et al. 1997; Parts et al. 2011).

Long-term selection creates two somewhat opposing forces for mapping. While each generation of selection increases the difference in QTL frequencies between divergently selected populations, recombination is expected to decrease the amount of linkage disequilibrium between markers and QTLs. Thus, unless the markers and QTLs are tightly linked, fairly rapid QTL allele-frequency changes (i.e., QTLs of modest to large effects) are required for a linked marker to show significant frequency changes (Lebowitz et al. 1987). Keightley and Bulfield (1993) showed that the marker and QTL must be closer than 20 cM for significant marker allele-frequency changes to be likely. One can use selective sweep theory (WL Chapter 8) to obtain more exact values.

Given the small population sizes of most selection experiments, drift alone is expected to change allele frequencies, motivating the need for tests for whether observed changes exceed those expected by chance events alone (reviewed in Chapter 9 of WL). The selection coefficient s on a QTL, which determines how quickly allele frequencies change (WL Equation 5.3), is a function of the QTL effect and the amount of selection on the character (WL Equation 5.21). Thus, s can provide information about the size of QTL effects. A maximum likelihood approach for estimating s was suggested by Keightley and Bulfield (1993) and Keightley et al. (1996). Here, computer simulations generate the distribution of allele-frequency change between two populations under divergent selection (assuming selection coefficient s) and drift (using an estimate of the effective population size). This procedure is repeated over a range of possible s values, and the value giving the highest probability of the observed change is taken as the ML estimator of s.

Marker alleles associated with body size have been found in several studies of divergently selected lines of mice (Garnett and Falconer 1975; Simpson et al. 1982; Gray and Tait 1993; Keightley and Bulfield 1993; Keightley et al. 1996). For example, after 22 generations of selection, Keightley and Bulfield (1993) found significant changes at four of 16 assorted marker loci (two coat color, two allozyme, 12 proviral inserts). This same set of crosses was further examined using 124 microsatellite loci by Keightley et al. (1996), who found 11 QTLs with estimated effects ranging from 0.17 to 0.28 phenotypic standard deviations. In their pioneering study in maize, Stuber at al. (1980) found that 8 of 20 allozyme markers showed significant differences for yield in divergently selected lines starting from F_2 crosses between inbred lines. To test whether these associations might simply be due to chance, Stuber et al. (1982) independently selected on the markers alone, finding a response for yield in the direction predicted by the original marker-trait associations. Other examples using allele frequency change to map QTLs in maize include Labate et al. (1999) and Coque and Gallais (2006) for yield, and Wisser et al. (2008) for disease resistance, while De Koeyer et al. (2001) examined yield in oats. Finally, an important special case of using allele-frequency change to map QTLs is in the search for longevity genes (Christensen et al. 2006; Murabito et al. 2012). Here, one contrasts marker allele frequencies between a control group and a group showing extreme longevity (such as a cohort of humans over 90 years old).

Example 17.9. Nuzhdin et al. (1993) used allele-frequency changes to detect markers associated with fitness in a cross of high and low fitness lines of *Drosophila melanogaster*. These authors used the presence/absence of the mobile genetic elements *mdg1* and *copia* as markers, finding 19 locations on chromosome 2 where the high and low lines differed in the presence/absence of inserted elements. An F_1 was backcrossed to the low line for three generations to generate a base population with high-line alleles at expected frequency $1/32$. This base population was then allowed to reproduce, with marker frequencies sampled after 11, 13, and 17 generations of natural selection. The frequencies of nine high-line markers (all located in a region around the centromere of chromosome 2) showed significantly higher values than expected by drift, suggesting linkage to one (or more) QTLs increasing fitness. The associated selection coefficients estimated by the ML approach of Keightley and Bulfield ranged from 0.3 to 0.7.

MARKER-BASED ANALYSIS USING INTROGRESSION LINES

NILs, HIFs, and STAIRS

The use of **introgression lines**, **ILs** (Eshed and Zamir 1995), has been proposed for both detection and fine mapping of QTLs. These are the extension of chromosome assays, but instead of whole chromosomes being replaced in an otherwise standardized background, much smaller chromosomal segments are substituted (from a **donor line**) into an otherwise fully homozygous background. Figure 17.8 shows the differences between two common IL population structures, NILs and HIFs. These differ in the uniformity the homozygous background, either being entirely from one parent (NILs) or a roughly equal mixture from both parents (HIFs).

Nearly isogenic (also **near-isogenic**) **lines**, **NILs**, are inbred lines containing one or more small regions of DNA from a **donor parent** in an otherwise standard background that is the same over all lines (Muehlbauer et al. 1988; Kaeppler et al. 1993; Monforte and Tanksley 2000). These are also referred to as **congenic strains** (Loosli et al. 1961) or **backcross inbred lines**, **BILs** (Wehrhahn and Allard 1965; Jeuken and Lindhout 2004; Kooke et al. 2012). NILs are constructed by first crossing a donor parent to an inbred line (the **recurrent parent**) to form an F_1. The resulting offspring are then backcrossed to the recurrent parent for several generations to generate a background consisting of almost entirely recurrent DNA. In species where selfing is allowed, the NIL is formed by at least one generation of selfing following the backcrossing. Where selfing is not possible, after the backcrossing is complete, individuals are repeatedly sib mated to form the final inbred line.

NILs differ subtly, but significantly, from recombinant inbred lines (RILs), see Figure 17.8. While both are inbred and usually start from F_1 parents, RILs are generated by repeated inbreeding of the descendant lines and hence contain about 50% donor DNA, as opposed

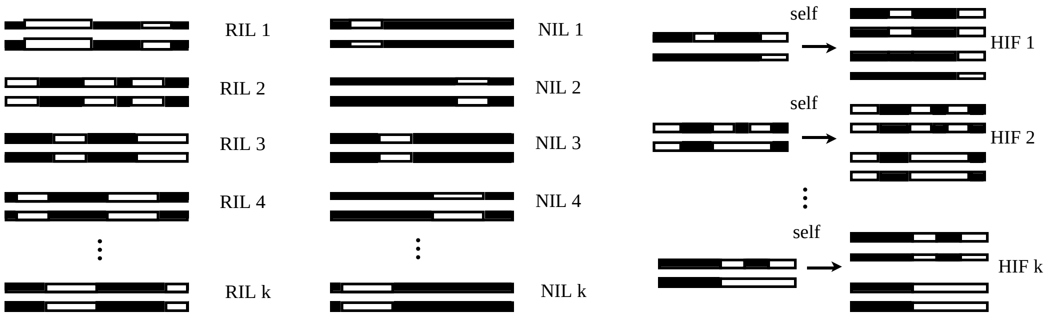


Figure 17.8 Different inbred-line population structures for mapping (RILs) and fine mapping (NILs, HIFs) QTLs. Solid and clear segments represent, respectively, segments of recurrent and donor DNA. **Left:** RILs are a collection of inbred lines that contain roughly equal amounts of random segments from the two parental lines. A contrast of any two RILs, or a RIL with either parental line, compares the cumulative impact from all regions at which the two lines differ. **Center:** NILs provide a much cleaner approach for fine mapping, as comparing a NIL against the recurrent parent contrasts only a single focal region. **Right:** Heterogeneous inbred families (HIFs) represent an intermediate strategy between RILs and NILs. Lines that are not yet fully inbred will still be segregating at a few loci (left side of the panel), with these regions becoming differentially fixed in the resulting family of inbred offspring from that individual (right side of the panel). In such a family, the backgrounds are identical except for the differential fixation of the segregating regions. Hence, within-family comparisons (e.g., two lines from HIF 1) results in a NIL-like contrast of the impact of the differentially fixed region. The key difference between NILs and HIFs is that a region of interest is always contrasted in the same background in NILs (the recurrent parent), while the background for the comparison of HIFs is a roughly equal mixture from both parents. Comparison of a region over a series of different HIFs provides a better measure of epistasis (assayed over multiple genetic backgrounds) than in the NIL setting (assayed over a single background).

to the very small fraction of donor DNA expected in NILs due to repeated backcrossing to the recurrent parent. In particular, the expected proportion, $p(b)$, of the donor genome left in a NIL following b generations of backcrossing and a final generation of selfing is

$$p(b) = (1/2)^{b+1} \tag{17.9}$$

Hill (1993) gives expressions for the variance of $p(b)$, which depends on number of chromosomes and their map distances. Typically, five to seven generations of backcrossing are performed, giving an expected proportion of donor genome of 1.2% to 0.5%. This repeated backcrossing scheme is the basis for the Wehrhahn-Allard estimator of gene number, discussed in Chapter 11.

Two different approaches of NIL formation have be used to map, and fine-map, QTLs: **random NILs** (backcrossing without selection) and **targeted NILs** (selection for either traits or specific markers during backcrossing). Markers greatly facilitate both approaches: (i) by characterizing the size and location of introgressed regions under either random or trait-selected NILs, and (ii) by marker-assisted selection for marker-targeted NILs. In the formation of random NILs, one simply backcrosses to the recurrent parent without any selection. The result is a collection of lines with random introgressed regions from the donor parent in an otherwise recurrent background. Because the introgressed segments represent a random sample of the donor genome, one can use a set of random NILs to map QTLs for essentially any trait (Examples 17.10 and 17.11).

The search for QTL regions typically proceeds by testing whether the trait value of a particular NIL is statistically different from that of the recurrent parent. If yes, then one or more QTLs for the trait are located in the regions where the NIL differs from the recurrent parent. The problem of jointly testing multiple NILs is often controlled for by using

Dunnett's test (Dunnett 1955, 1964) for multiple comparisons against a single control (the recurrent parent). Other methods of analysis of NILs were explored by Falke and Frisch (2011) and Mahone et al. (2012), who found that a linear model approach (treating each introgressed segment as an effect to be estimated), combined with Holm's extension of the Bonferroni correction for multiple testing (Appendix 6), was generally more powerful than Dunnett's test.

While both RILs and random NILs can be used as resource populations for QTL mapping, they represent slightly different strategies. The relative efficiency of these two approaches has been examined (both theoretically and empirically) by a several authors (Kaeppler 1997; Keurentjes et al. 2007; Szalma et al. 2007). Generally speaking, RILs have greater power of QTL detection, while NILs offer more precise estimates and the ability to detect QTLs of smaller effect. When epistasis is widespread, RILs offer the advantage over NILs in that a genomic region is tested over several backgrounds versus over a single background in NILs. While using NILs reduces the background genetic noise and thus increases power, it can also provide a misleading picture of the genetic architecture if variation in the genetic background has a significant effect on the target region.

With targeted NILs, one backcrosses to the recurrent parent, but selects for either a specific trait (Example 17.12) or a particular marker (next section), such as that around segment suspected of containing one, or more, QTLs. While targeted NILs allow for fine mapping, unlike randomly formed lines, these lines are not that useful for mapping QTLs for arbitrary traits, because the introgressed segments do not represent a random sample of donor genome.

Example 17.10. Eshed and Zamir (1995) formed 50 NILs from a cross between the cultivated tomato (*Lycopersicon esculentum*) and a wild relative (*L. pennellii*). Based on analysis with 375 markers, each line contained a single RFLP-defined fragment from *L. pennellii*, averaging around 33 cM (or roughly 3% donor DNA). Comparing the mean value of each NIL against a standard, the authors found 23 QTLs for fruit soluble-solids content and 18 QTLs for fruit mass. By comparison, in an analysis of marker-trait associations in RILs, Goldman et al. (1995) found 7 and 13 QTLs, respectively. Eshed and Zamir subjected two regions to much finer scale mapping, using markers to select for recombinants within these regions. Upon finer analysis, a 55 cM region influencing fruit size was shown to contain at least three separate QTLs. A 37 cM region showing heterosis for soluble-solids yield was shown to result from associative overdominance (Chapters 12 and 13). This latter region was subdivided by recombination into a partially dominant QTL that increases yield and a linked recessive that reduces yield.

Example 17.11. Obando et al. (2008) formed a set of 97 melon (*Cucumis melo*) NILs from a cross between a Spanish (recurrent parent) and a Korean (donor parent) accession. Microsatellite markers were used to characterize the size and locations of the Korean introgressed segments, with the set of random NILs covering roughly 85% of the Korean genome, with an average introgression size of around 41cM (roughly 3% of the Korean genome). Ten replicates of each NIL, along with 50 replicates of the Spanish line and five replicates of the Korean line, were grown in a completely randomized design (Appendix 9). Roughly thirty morphological and fruit-related traits were scored, and the mean value for each trait for a given NIL replicate was compared against the mean for the control Spanish line using Dunnett's test. Based on this approach, a total of 134 QTLs were detected, 52 for morphological traits, 69 for fruit color traits, and 23 for flavor traits. Every introgressed region showed a QTL for at least one trait, with most have two or more QTLs (an average of 3.5 QTL per genomic region, i.e., each region impacted, on average, 3.5 of the scored traits).

Example 17.12. Targeted NILs are constructed by a select-and-backcross procedure, and hence are not random with regard to the retained donor DNA (Wright 1952; Hill 1998a). "Exotic" sorghum is tall and has a short-daylength requirement for flowering, while commercial sorghum cultivars are short (for easier harvesting) and day-length insensitive. Exotic strains are "converted" to commercial cultivars by first crossing to a standard short, daylength-neutral donor parent and continually backcrossing each generation to the exotic parent while selecting for short height and daylength neutrality. The resulting converted strains are expected to contain mostly exotic DNA, but are daylength neutral and short, due to the retention (by selection) of donor QTL alleles influencing these traits. Hence, one approach to mapping QTLs for these traits is to search for regions that are retained in the converted derivatives.

This approach was used as a check of conventional QTL mapping by Lin et al. (1995), who examined 71 markers in nine exotic accessions of *Sorghum bicolor* and their converted derivatives. The authors compared the positions of retained donor regions in these accessions with the positions of QTLs for height and daylength sensitivity mapped in a previous experiment using the F_2 progeny in a cross of *S. bicolor* $\times$ *S. propinquum*. Seven of the nine QTLs (six for height, three for daylength) detected using the F_2 cross coincided with regions of donor parent DNA retained in the converted strains despite several generations of backcrossing. One donor region was seen in all nine derivatives, and in the F_2 crosses this region accounted for 55% of the variation in height and 86% of the variation in daylength sensitivity. In contrast, three converted regions (found in at least one line) did not coincide with mapped QTLs. Whether these regions indeed contain QTLs for height or daylength could be examined by crossing the converted strains with their parental exotics and looking for associations between the retained donor markers and trait values.

A second approach for generating ILs is through the use of **heterogeneous inbred families**, or **HIFs** (Haley et al. 1994; Tuinstra et al. 1997). As shown in Figure 17.8, this approach exploits the fact that while a line undergoing inbreeding is isogenic for much of its genome (having become homozygous), there are regions where it is still segregating alternative segments (residual heterozygosity). Some of the selfed offspring from such parents will be fixed for different segments, generating NILs. For example, after five generations of selfing, roughly 94% of the loci in a line are fixed as homozygotes, while 6% are heterozygotes, resulting in contrasting members of a HIF family being differentially fixed at these segregating regions. The major difference between NILs and HIFs is the nature of the background homozygous genome, which is entirely from the recurrent parent in NILs, but roughly half recurrent and half donor in HIFs (Figure 17.8). The strength of a NIL is that any additional noise from background variation is removed, increasing power. The weakness of NILs is that if epistasis is present, the donor and recurrent regions are only contrasted in the recurrent background, while under HIFs, the contrast is across a variety of different backgrounds.

A final structured inbred population that is occasionally used for QTL fine mapping are **stepped aligned inbred recombinant strains**, or **STAIRS** (Koumproglou et al. 2002; Perera et al. 2006). This is approach represents an intermediate strategy between chromosomal substitution lines and NILs. Here, one generates a series of single recombinants for each chromosome, with donor DNA distant, and recurrent DNA proximal, to the recombination breakpoint. This creates a nested series for each chromosomes that contain successively longer tracts of donor population DNA. Koumproglou et al. generated a series of such lines in *Arabidopis* that allowed for mapping to regions of around 0.5cM by a simple comparison between a few STAIRS lines for the chromosome of interest. As with (random) NILs and HIFs, the construction of a collection of STAIRS requires a considerable amount of time and effort, but the resulting resource makes a powerful tool for mapping any trait showing variation between the chosen recurrent and donor lines.

Marker-based Introgressions

The formation of targeted NILs requires constant backcrossing to the recurrent parent coupled with selection for donor DNA. Historically, selection was phenotypic, retaining

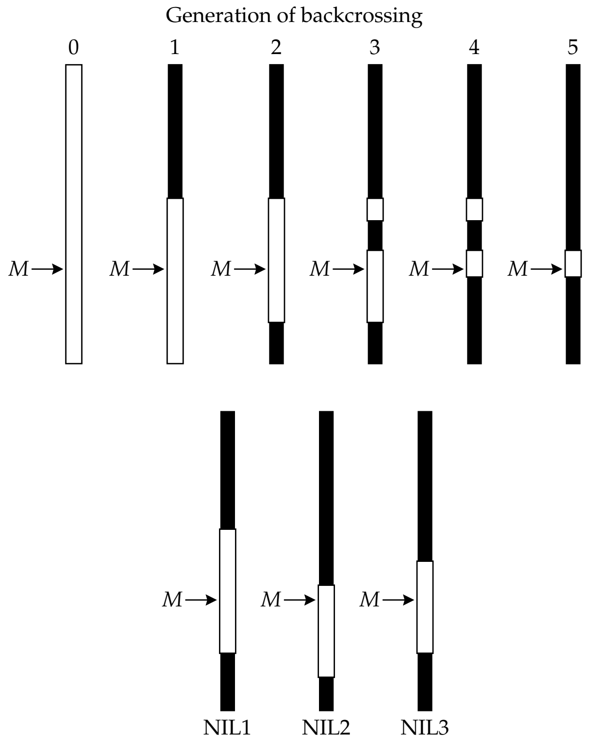


Figure 17.9 **Top:** Construction of a nearly isogenic line (NIL) containing a region of interest, obtained by selecting on a single marker *M* while backcrossing to a standard stock (the recurrent parent). Because the final lines are constructed by selfing, we follow a single (haploid) chromosome, as the NIL has two identical copies of this chromosome. Clear regions are from the donor parent, solid from the recurrent parent. **Bottom:** If several lines are independently selected for the marker, the resulting NILs will contain different retained regions around *M*. With a dense marker map, crosses between such lines allow for finer mapping of the region around *M*.

backcrossed individuals showing the trait of interest such as disease resistance or a small set of desirable characters (e.g., Example 17.12). In the marker era, one can instead focus on creating NILs around a region of interest that is anchored by one (or more) markers. Such **marker-assisted backcrossing** allows the introgression of specific regions (such as one containing previously detected QTLs) into an otherwise isogenic background (e.g., Young et al. 1988; Tanksley et al. 1989; Paterson et al. 1991; Dudley 1993). Independent NILs formed while selecting on the same marker are expected to differ in the regions retained around the marker (Figure 17.9), and crossing such NILs allows for finer scale mapping. This approach has been used to fine map QTLs (mainly for genes of moderate to large effects, often resistance genes) in maize (Bentolila et al. 1991; Koester et al. 1993; Touzet et al. 1995; Liu et al. 2012), rice (Yu et al. 1991; Ronald et al. 1992; Huang et al. 2013), lettuce (Paran et al. 1991; Den Boer et al. 2013), wheat (Hartl et al. 1993; Schachermayr et al. 1994, 1995; Zheng et al. 2015), barley (Schüller et al. 1992; Hinze et al. 1991; Vu et al. 2010), oats (Penner et al. 1993), soybeans (Kim et al. 2010), and tomatoes (Osborn et al. 1987; Tanksley and Hewitt 1988; Sarfatti et al. 1989; Paterson et al. 1990; Kinkade and Foolad 2013).

As noted by Hospital et al. (1992), marker-assisted selection during the creation of introgression lines has two goals: (i) **foreground selection** to retain a desired donor segment in the face of continual backcrossing to the recurrent line, and (ii) **background selection** to accelerate the removal of donor DNA from the rest of the genome. As the available marker density increased, foreground selection progressed from using single to flanking markers, while the availability of markers outside of the donor region facilitated background selection throughout the rest of the genome. The combination of both marker-assisted foreground and background selection to construct NILs has been called **speed congenics** (Markel et al. 1997; Wakeland et al. 1997; Visscher 1999; Wong 2002).

We start by considering foreground selection using only a single marker in the donor region of interest. Selecting for the retention of a donor marker while backcrossing to the recurrent parent yields a NIL with a block of donor DNA surrounding the marker that can be rather large. Linked undesirable genes can be dragged along with the marker, a phenomenon referred to as **linkage drag** by plant breeders (Brinkman and Frey 1977), and as **Hill-Roberston effects** by evolutionary biologists (Hill and Robertson 1966; WL Chapter 8). Conversely, a single marker may remain associated with several favorable linked QTLs within the introgressed region. The amount of donor DNA present following marker-assisted foreground selection has been examined by several authors (Bartlett and Haldane 1935; Hanson 1959a–1959d; Stam and Zeven 1981; Naveira and Barbadilla 1992; Frisch and Melchinger 2001). The donor DNA remaining in the NILs can either reside on the chromosome under marker selection or on the other (unmarked) chromosomes. We consider these two sources separately.

Suppose there are n chromosomes and a total genome length of $\mathcal{L} = \ell_M + \sum \ell_i$, with ℓ_M being the length of the marker chromosome and ℓ_i being the length of the *i*th unmarked chromosome (all lengths in Morgans). Consider the DNA retained around the marker, and suppose that the selected marker is in the center of the chromosome. Letting $t = b + 1$, the proportion $p_M(b)$ of donor DNA on this marker chromosome after b generations of backcrossing and a final generation of selfing has expected value

$$E[\, p_M(b)\,] = 2\left(\frac{1 - e^{-t\ell_M/2}}{t\ell_M}\right) \tag{17.10a}$$

(Hanson 1959b), while the variance is

$$\sigma^2[\, p_M(b)\,] = \frac{2}{(t\ell_M)^2}\left[1 - \left(t\ell_M + e^{-t\ell_M/2}\right)e^{-t\ell_M/2}\right] \tag{17.10b}$$

(Naveira and Barbadilla 1992). For $t\ell_M \gg 1$, these are approximately

$$E[\, p_M(b)\,] \simeq \frac{2}{t\ell_M}, \qquad \sigma^2[\, p_M(b)\,] \simeq \frac{2}{(t\ell_M)^2} \tag{17.10c}$$

Since these equations refer to the proportion of this chromosome retained, the corresponding chromosome lengths retained are given by $\ell_M \cdot p_M(b)$, yielding a mean length ± SD (in Morgans) of $2/t \pm \sqrt{2}/t$ for $t\,\ell_M \gg 1$.

The restriction of the marker being in the center of the chromosome was removed by Stam and Zeven (1981), who showed that using Equations 17.10a–17.10c for a randomly placed marker generally introduces only a small error. Of broader concern is that these equations consider only the *segment around the marker*, while the marked chromosome can also retain blocks of donor DNA elsewhere. Stam and Zeven show that these two effects (noncentrality of the marker and blocks of donor DNA retained outside of the marker) largely cancel, leaving Hanson's original result as a very good approximation.

Turning now to the untagged chromosomes, let $p_{U_i}(b)$ denote the proportion of donor DNA on an untagged chromosome of length ℓ_i. The mean and variance are given by

$$E[\, p_{U_i}(b)\,] = \frac{1}{2^t} \tag{17.11a}$$

$$\sigma^2[\, p_{U_i}(b)\,] = 2\left(\frac{1 - e^{-t\ell_i/2}}{t\,\ell_i\, 2^t}\right) - \frac{1}{2^{2t}} \tag{17.11b}$$

as shown by Stam and Zeven (1981). The expected total length of donor DNA retained over all untagged chromosomes thus becomes

$$\sum_{i=1}^{n-1} \ell_i \, E[\, p_{U_i}(b)\,] = \frac{1}{2^t} \sum_{i=1}^{n-1} \ell_i = \frac{\mathcal{L} - \ell_M}{2^t} \tag{17.12a}$$

with variance

$$\sum_{i=1}^{n-1} \ell_i^2 \cdot \sigma^2[\, p_{U_i}(b)\,] \tag{17.12b}$$

Example 17.13. Suppose there are $n = 15$ chromosomes, each of length 75 cM, and that we select for retention of a marker located in the middle of chromosome 1 during five generations of backcrossing followed by a generation of selfing. Applying Equation 17.10a, the expected fraction of donor DNA on chromosome 1 is

$$E[\, p_M(5)\,] = 2 \left(\frac{1 - e^{-6\,(0.75/2)}}{0.75 \cdot 6} \right) \simeq 0.40$$

or, in units of chromosome length, 0.4 · 75 cM = 30 cM. Equation 17.10b gives the standard deviation of the expected proportion retained as 0.23, or 17.25 cM. Likewise, applying Equations 17.11a and 17.11b, the proportion of donor DNA on any untagged chromosome is 0.016 ± 0.077, or 1.2 ± 5.8 cM. The fraction of the total expected donor DNA in the NIL contributed by the marked chromosome is 30/(30 + 14 · 1.2) = 0.64. Values for other generations are

	Lengths of Introgressed Segment (cM)		% from
b	Marker chromosome	All unmarked	Marked Chromosome
2	45.0 ± 19.1	131.25 ± 255.9	25.6
7	23.8 ± 14.8	4.10 ± 36.7	85.3
10	17.9 ± 12.0	0.51 ± 11.3	97.3
12	15.3 ± 10.5	0.13 ± 5.2	99.2
15	12.5 ± 8.7	0.02 ± 1.7	99.9

Finally, we note that a NIL line created by selecting for retention of a single marker thus contains ***significantly more donor DNA than a NIL created without any selection***. Here 1,125 cM is the total map length, giving the expected total length of donor DNA in an NIL formed without selection as $1125/2^{b+1}$, or 17.58, 4.39, and 0.55 cM after 5, 7, and 10 generations of backcrossing, respectively.

If donor and recurrent-parent DNAs are sufficiently divergent, recombination between donor and recurrent chromosomes can be greatly reduced, resulting in a much longer segment of donor DNA being retained than expected (e.g., Young and Tanksley 1989a; Paterson et al. 1990). In these cases, the above expressions may give rather serious underestimates. For example, Young and Tanksley (1989a) examined the length of donor DNA introgressed by selecting for a donor marker (resistance to tobacco mosaic virus) in a cross between two species of tomatoes. Even after 11 generations of backcrossing, the length of the introgressed segment in one line was over 51 cM, as compared to the theoretical expectation of 16.7 ± 11.8 cM from Equation 17.10c.

Selection using *multiple*, instead of *single*, markers can greatly accelerate the formation of NILs with very short introgressed segments. First, foreground selection using *flanking markers* around a region of interest can greatly reduce the size of the introgressed region (Young and Tanksley 1989a, 1989b). Second, background selection for markers from the recurrent parent on other chromosomes can accelerate the loss of donor DNA outside of the introgressed region. We consider these two strategies in turn.

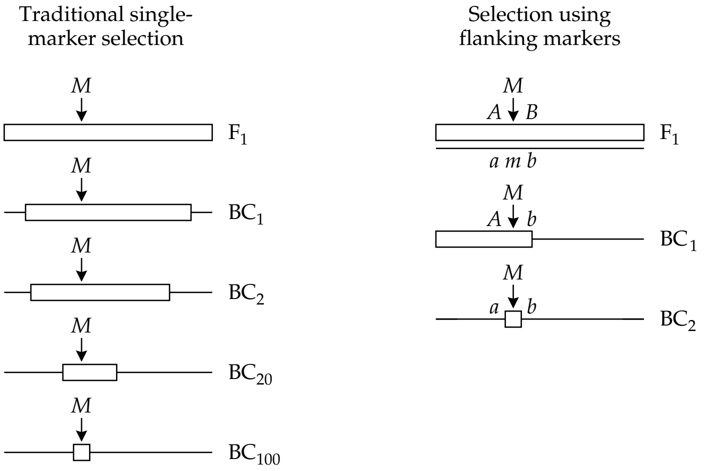

Traditional single-marker selection
Selection using flanking markers
M
F1
BC1
BC2
BC20
BC100
A B
a m b
A b
a b

Figure 17.10 Selecting on flanking markers can greatly accelerate the elimination of donor DNA around a selected donor marker *M*. **Left:** Reduction of donor material around a single selected marker leads to an expected value of around 2 cM following 100 generations of selection and backcrossing to the recurrent parent. **Right:** Two generations of selection using a dense marker map can also reduce the introgressed segment to under 2 cM. First, choose two flanking markers (*a* and *b*) from the recurrent parent that are both within 1 cM of *M*, with the donor and recurrent genotypes being *AMB* and *amb*, respectively. In generation one, we select an individual that is either *AMb* or *aMB*. Using this individual as the parent for the second generation, we then select for the recombinant on the opposite side, giving an *aMb* individual with an introgressed region containing *M* that is less than 2 cM. (After Tanksley et al. 1989.)

We consider two-marker foreground selection first. Equation 17.10c shows that 99 backcross generations are required to reduce the average length of a marker-selected introgressed region to 2 cM. However, with a dense marker map, such a region can be obtained in one or two generations by using markers that flank the region of interest (Figure 17.10). If we can screen a very large number of individuals, this can be done in a single generation by looking for a double recombinant. Using flanking markers 1 cM on either side of the region of interest, the probability of a double recombinant (assuming no interference) is 0.01^2, or about one out of every 10,000 progeny. If interference is significant, which is likely for tightly linked markers, far more progeny may need to be scored. This problem can be avoided by screening for two generations, with first-generation individuals showing a single recombinant (on either side of the marker) being used as parents for the second generation.

Example 17.14. Suppose the region of interest is flanked by markers, each at recombination frequency c from the locus of interest. To have a 95% probability of recovering a double recombinant in a single generation (assuming no interference), one must examine roughly $3/c^2$ individuals, or 30,000 individuals for c = 0.01 (a 2 cM interval). This follows since the probability of no double recombinants among n sampled chromosomes is $(1-c^2)^n$, giving the probability of at least one double as $1 - (1-c^2)^n$. Setting this equal to 0.95 and solving for n gives $n = -\ln(0.05)/\ln(1-c^2) \simeq 3/c^2$. How much of a savings can the two-generation method offer?

If n_1 first-generation individuals are screened, the probability that at least one of them is a single recombinant is $1-(1-2c)^{n_1}$ (with $2c$ being the probability of recombination on either side of the marker, $1-2c$ is the probability that a chromosome is nonrecombinant with respect to both flanking markers). Using a single recombinant to form the next generation, the probability of a single recombination event on the opposite side of the region is $1-(1-c)^{n_2}$. The probability of obtaining a double recombinant after two generations is the probability of a single recombinant in the first generation times the probability of a single recombinant in the second (this result is not impacted by interference). Thus, to have an overall probability of

95%, we need a probability of 97.5% for each of the two events, as $0.975^2 = 0.95$. The sample size required to have a 97.5% chance of a single recombinant (on either side of the region) in the first generation is obtained by solving $0.975 = 1 - (1-2c)^{n_1}$, yielding $n_1 \simeq 1.85/c$. Likewise, solving $0.975 = 1 - (1-c)^{n_2}$ gives $n_2 \simeq 3.7/c = 2\,n_1$. Using this two-generation approach, to have a 95% probability of recovering a double recombinant, a total of $n_1 + n_2 = 5.55/c$ individuals must be scored. For $c = 0.01$, this is 555 individuals (185 in the first generation, 370 in the second), only two percent of the 30,000 required for the single-generation method. Frisch et al. (1999a) provide a more in-depth analysis of sample size and optimal placement of the flanking markers.

Turning now to background selection, Hillel et al. (1990; 1993) first proposed that selection for recurrent markers (in their case in chickens) could greatly hasten the removal of background donor DNA, although they were overly optimistic about its advantage (Hospital et al. 1992; Groen and Smith 1995; Visscher 1999). In actuality, marker-assisted background selection results in around a two to three generation reduction in the amount of time to reach a desired fraction of recurrent DNA compared to backcrossing without selection (Hospital et al. 1992; Visscher 1996, 1999; Visscher et al. 1996; Frisch et al. 1999b; Ribaut et al. 2002; Stam 2003).

The optimal number, and placement, of markers for efficient background selection has been examined by a number of authors, with the general conclusion of between two and four markers per Morgan of chromosome length, ideally positioned (for metacentric chromosomes) around 20-30 cM from the telomeres (Hospital et al. 1992; Visscher 1996; Visscher et al. 1996; Frisch et al. 1999b; Servin and Hospital 2002; Frisch and Melchinger 2005). There is a rapidly diminishing return from adding new markers per chromosome (Herzog and Frisch 2011), as there are simply not enough recombinations within a sample to exploit this addition fine-mapping information. Hence, the limiting factor is usually the number of recombinations (i.e., number of scored individuals), rather than the ability to finely map their breakpoints (number of markers). This is a recurrent theme in QTL mapping: the number of individuals sampled, rather than the number of markers scored, is generally the limiting factor. ***Far more information is gained in a study by adding more individuals than by increasing the marker density*** (e.g., Frisch and Melchinger 2005). For example, as noted by Kump et al. (2010), the usual fine-mapping strategy of first localization of a QTL using RILs and the subsequent fine-mapping by creating NILs for the target region can be significantly improved by combining information from both data sets, as more potential recombinants are scored.

Finally, one can combine reproductive strategies that significantly speed up generation time with marker-assisted selection, an approach termed **supersonic congenics** (Behringer 1997). For many vertebrates, **superovulation and embryo transfer** (hormone treatment that allows for the harvesting of eggs from prepubertal females) dramatically shortens generation times (e.g., Landa et al. 2009). In plants, similar gains can be made using **speed breeding** protocols that use extreme photoperiods and plant hormones to speed development (Ghosh et al. 2018; Watson et al. 2018; Chiurugwi et al. 2019). These rapid advancement-of-generations schemes do not typically allow for the scoring of normal phenotypes, but are suitable to maker-assisted selection.

FINE MAPPING MAJOR GENES USING POPULATION-LEVEL DISEQUILIBRIUM

Limitations for Pedigree-based Fine Mapping

The typical level of mapping resolution for a major gene using the linkage information in a collection of (human) pedigrees is rarely below 1–2 cM (Lange et al. 1985; Bodmer 1986; Boehnke 1994). To a rough approximation, this typically corresponds to (on average) a region of a few megabases in size (Table 17.2), and hence a few hundred potential candidate genes. The limiting factor is the expected number of scoreable recombinants,

$$\lambda = N n_o c \tag{17.13}$$

in the pedigree. Here N is the number of **informative parents** (individuals that are double heterozygotes for both the marker and the major gene, allowing for recombinants to be detected; Chapter 19), n_0 is the average number of offspring for each such parent, and c the marker-gene recombination frequency. Outside of controlled crosses (Chapter 18), usually only a fraction of parents are informative (Chapter 19), so that N can be considerably smaller than the actual number of individuals in the pedigree. From Equation 17.13, a marker-gene map distance of 1 cM ($c = 0.01$) requires 100 offspring from informative parents ($Nn_o = 100$) to have an expected value of one recombinant ($\lambda = 1$). The chance of not seeing any recombinants is $(1-c)^{Nn_o} \simeq e^{-\lambda}$, or $e^{-1} = 0.37$. More generally, the probability of seeing exactly k recombinants approximately follows for the Poisson distribution, $\lambda^k \exp(-\lambda)/k!$ (Equation 2.21a). Considering the first three terms ($k = 0, 1, 2$) for a set of pedigrees containing 200 total offspring from informative parents ($Nn_o = 200$), the probability of seeing two or fewer recombinants (for $c = 0.1$) is almost 60% (0.586).

Linkage-disequilibrium Mapping (LDM) Using Historical Recombinants

One source of additional linkage information is from **historical recombination**: the descendants of a single ancestral haplotype. Here the expected number of recombinants given a last common ancestor of t generators ago is $2tc$. Hence, two descendant haplotypes with a last common ancestor 100 generations ago have experienced the same expected number of recombination events as 200 offspring from informative parents in a pedigree. This introduces the idea of exploiting population-level linkage disequilibrium (LD) for gene mapping. Our focus here is for the very special case of mapping a *major gene*, an approach we denote as **linkage-disequilibrium mapping** (**LDM**). Chapter 20 examines the much wider use of LD in GWAS approaches for mapping genes of arbitrary effect. This latter approach requires a very dense marker map, but (given sufficient marker density) is applicable to many, if not most, natural populations (Chapter 20).

Here, our focus is a bit different, assuming a lower density of markers with the goal to estimate a map distance between a closely linked marker and the target gene. The basic idea was suggested by Bodmer (1986, cited by Hill and Weir 1994), wherein once the rough (5–10 cM) position of a gene is found,

> ". . . it should be possible to saturate the relevant region with further polymorphic markers and look for the ones that have a population association with the trait, rather than increasing the numbers of families analyzed to search for closer linkage. This is the most efficient way of finding closely linked markers, since [human] family data are very inefficient at distinguishing between small recombination fractions such as 0.5% versus 5.0%."

This requires a setting in which the LD around the major gene is sufficiently large to encompass a marker that may be some distance away. Such larger regions of LD can occur in expanding populations, and geneticists have sought out isolated human populations that appear to have recently undergone such expansions for LD mapping of disease genes. The ideal situation occurs when most of the disease alleles descend from a single ancestral mutation, so that all current copies have maintained some of the ancestral marker haplotype. The age of the ancestral mutation from which all current copies derive can neither be too young nor too old. It must be sufficiently old that recombination has reduced the expected size of the retained region to one sufficiently small for fine mapping. However, if the mutation is too old, this region will be too small for analysis (with only modest marker density, no variation may be present around the gene). These conditions are often satisfied in expanding populations that can be traced back to a small number of recent founders. An example is the current Finnish population, which arose from a small founder group about 2000 years ago (reviewed in de la Chapelle 1993; de la Chapelle and Wright 1998; Peltonen et al. 1999), and other such human populations are discussed by Jorde (1995) and Peltonen et al. (2000). In a rapidly expanding population, all copies of a mutant disease allele often trace back to a single copy present in the founding population or originating shortly after the population expansion began.

Motivated by the Luria-Delbrück theory for estimating mutation rates in exponentially growing populations of bacteria (Luria and Delbrück 1943; Sarkar 1991); Hästbacka et al. (1992) and Lehesjoki et al. (1993) suggested that the same logic can be used to estimate recombination frequencies from population disequilibrium information. The simplest approach proceeds as follows. Suppose a disease allele (of major effect) was either present as a single copy (and hence associated with a single chromosomal haplotype) in the founder population or arose by mutation very shortly after the population was formed. Further assume that there is no **allelic heterogeneity**, so that all disease-causing alleles in the population descend directly from the original mutation, and consider a marker locus tightly linked to the disease locus. The probability that a disease-bearing chromosome has not experienced recombination between the **disease susceptibility** (**DS**) gene and marker locus after t generations is just $(1-c)^t \simeq e^{-c\,t}$, where c is the marker-DS recombination frequency. Suppose the disease is predominantly associated with a particular haplotype, which presumably represents the ancestral haplotype on which the DS mutant arose. Equating the probability of no recombination to the observed proportion, π, of disease-bearing chromosomes with this predominant haplotype gives $\pi = (1-c)^t$, where t is the age of the mutation or the age of the founding population (whichever is more recent). Hence, one estimate of the recombination frequency is

$$c = 1 - \pi^{1/t} \tag{17.14}$$

Note the key assumption throughout (that greatly limits the generality of this approach): that ***the phenotype unambiguously identifies the presence of a DS allele***. This restriction may be true for some major genes, but not for general genes underlying a complex trait. Hence, this approach is essentially restricted to Mendelizing factors (Kaplan et al. 1995), while with much denser marker maps (and much large population sample sizes), GWAS offers a more general, and more powerful, approach (Chapter 20).

Example 17.15. Hästbacka et al. (1992) examined the gene for diastrophic dysplasis (DTD), an autosomal recessive disease, in Finland. A total of 18 **multiplex** families (showing two or more affected individuals) allowed the gene to be localized to within 1.6 cM from a marker locus (*CSF1R*) using standard pedigree methods. To increase the resolution using pedigree methods requires significantly more multiplex families. Given the excellent public health system in Finland, it was likely that the investigators had already sampled most of the existing families. As a result, the authors turned to LD mapping.

While only multiplex families provide information under standard mapping procedures, this is not the case with LD mapping wherein *single affected individuals* (which each contain two DTD-bearing chromosomes) can provide further information. Using LD mapping allowed the sample size to increase by 59 in this study. A number of marker loci were examined, with the *CSF1R* locus showing the most striking correlation with DTD. The investigators were able to unambiguously determine the haplotypes of 152 DTD-bearing chromosomes and 123 normal chromosomes for the sampled individuals. Four alleles (haplotypes) of the *CSF1R* marker gene were detected. The frequencies for these alleles among normal and DTD chromosomes were as follows:

	Chromosome type			
CSF1R Allele	Normal		DTD	
1-1	4	3.3%	144	94.7%
1-2	28	22.7%	1	0.7%
2-1	7	5.7%	0	0%
2-2	84	68.3%	7	4.6%

Given that the majority of DTD-bearing chromosomes were associated with the rare 1-1 allele (present in only 3.3% of normal chromosomes), the authors suggested that all DTD-bearing chromosomes in the sample descended from a single ancestor carrying allele 1-1. Since 95% of all present DTD-bearing chromosomes contain this allele, $\pi = 0.95$. The current Finnish population traces back to around 2000 years to a small group of founders, which underwent around $t = 100$ generations of growth. Using these estimates of π and t, Equation 17.14 gives an estimated recombination frequency between the *CSF1R* gene and the DTD gene as $c = 1 - (0.95)^{1/100} \simeq 0.00051$. Thus, the two genes were estimated to be separated by 0.05 cM, or about 50 kb (using the rough rule for humans that 1 cM = 10^6 bp). Subsequent cloning of this gene by Hästbacka et al. (1994) showed it be to 70 kb proximal to the *CSF1R* marker locus. Thus, LD mapping increased precision by about 34-fold over that possible using segregation within pedigrees (0.05 cM vs. 1.6 cM).

The DTD example is exceptional in that the disease allele was initially associated with a rare haplotype. Since the majority of current disease-bearing chromosomes have this haplotype, this suggested that the sample of disease alleles was unlikely to contain a significant fraction of new mutants. Lehesjoki et al. (1993) offered a modification when the predominant marker haplotype among disease-bearing chromosomes is also a common haplotype for normal chromosomes. Let p_{DS} denote the frequency of the most common haplotype of disease-bearing chromosomes and p_n denote the frequency of this haplotype among normal chromosomes. Let π represent the fraction of disease-bearing chromosomes descended from the common ancestor that have not undergone any recombination between the marker and QTL, and α be the proportion of all disease alleles descending from the founder copy (as opposed to being new mutants). We can then divide disease-bearing chromosomes into two classes. First, a fraction $\alpha\,\pi$ of these chromosomes have a disease allele that both traces back to the founder copy (α) and does so through chromosomes that have not undergone recombination between the marker and disease loci (π). Second, a fraction $(1-\alpha\,\pi)$ are either new mutants (which arose on some random haplotype) or a product of recombination between the founder copy and a random haplotype. In either event, the chance that such individuals have the predominant haplotype is just $(1-\alpha\,\pi)\,p_n$. Putting these two results together gives

$$p_{DS} = \alpha\,\pi + (1-\alpha\,\pi)\,p_n \tag{17.15a}$$

Rearranging Equation 17.15a and solving, we find that

$$\delta_p = \alpha\,\pi = \frac{p_{DS} - p_n}{1 - p_n} \tag{17.15b}$$

is a measure of the extent to which disease chromosomes are restricted to the predominant haplotype (Bengtsson and Thomsom 1981). Recalling Equation 17.14,

$$\pi = (1-c)^t = \frac{\delta_p}{\alpha} \tag{17.16}$$

giving an estimate of the recombination frequency as

$$c = 1 - \left(\frac{\delta_p}{\alpha}\right)^{1/t} \tag{17.17}$$

Note that for the DTD data in Example 17.15, $p_n \simeq 0$, so that $\delta_p \simeq p_{DS}$, and we assumed that all disease alleles descended from the founder copy ($\alpha = 1$), so that $c \simeq 1 - (p_{DS})^{1/t}$. Note also that if c is known, one can solve instead for t, the age of the mutation. Risch et al. (1995) used this approach to date the appearance of a mutation causing idiopathic torsion dystonia (ITD) disease in Ashkenazi Jews at about 350 years (WL Chapter 9 discusses other approaches for estimating the age of alleles).

Hästbacka et al. (1992) proposed simple confidence intervals for estimates of c given by Equations 17.14 and 17.17, but Kaplan and Weir (1995) showed these are downwardly biased, leading to intervals that are too narrow. de la Chapelle and Wright (1998) suggested an improved approximation that incorporates sampling error in the estimation of π. However, maximum likelihood methods (Appendix 4) generally provide better approximations of the confidence intervals (Kaplan et al. 1995; Terwilliger 1995; Xiong and Guo 1997).

The fraction α of all disease alleles that are direct descendants from the ancestral copy (as opposed to being new mutants) can be obtained as follows. Assuming that selection is weak on carriers (heterozygotes), new mutants that have arisen *after* the initial appearance of the original DS allele in the founding population should comprise a fraction $1-(1-\mu)^t \simeq \mu\, t$ of all chromosomes, as $(1-\mu)^t$ is the probability that no mutations have occurred. Here μ is the mutation rate at which new disease-causing alleles appear and t is the number of generations since the original mutation arose or since the founding of the population, whichever is more recent. If q is the frequency of disease alleles, then $\mu\, t/q$ is the expected fraction of disease-carrying chromosomes due to new mutants. Hence, the expected fraction of all disease-carrying chromosomes that descend from the original mutation is

$$\alpha = \left(1 - \frac{\mu\, t}{q}\right) \tag{17.18}$$

which, when applied to Equation 17.17, yields

$$c = 1 - \left(\frac{\delta_p}{1-(\mu\, t/q)}\right)^{1/t} \tag{17.19}$$

The basic LDM idea of considering the fraction of DS-bearing haplotypes than contain a specific marker allele has been generalized to allow for both multiple alleles and multiple loci (Kaplan et al. 1995; Devlin and Risch 1995; Risch et al. 1995; Terwilliger 1995; Xiong and Guo 1997; Meuwissen and Goddard 2000; Toivonen et al. 2000; Liu et al. 2001; Morris et al. 2002; Schaid 2004; Zöllner and Pritchard 2005). When multiple markers are used, this leads to tests of the ordering of the markers and the DS gene.

Example 17.16. Sulisalo et al. (1994) examined the major gene for cartilage-hair hypoplasia (CHH), an autosomal recessive disease, in the Finnish population. As in Example 17.15, pedigree information allowed unambiguous determination the haplotypes associated with most CCH-bearing chromosomes. The authors observed that 85% of these contain a particular allele at marker *D95163*, while only 41% of non-CHH chromosomes carry this allele. Hence,

$$\delta_p = \frac{p_{DS} - p_n}{1 - p_n} = \frac{0.85 - 0.41}{1 - 0.41} = 0.75$$

implying

$$c = 1 - \left(\frac{\delta_p}{\alpha}\right)^{1/t} = 1 - \left(\frac{0.75}{\alpha}\right)^{1/t}$$

To estimate the fraction α of all current CHH alleles that directly trace back to the ancestral copy, first note that the frequency of CHH alleles in Finland is estimated to be 0.0066. Assuming $t = 100$ and $\mu = 1 \times 10^{-5}$,

$$\alpha = \left(1 - \frac{\mu\, t}{q}\right) = \left(1 - \frac{100 \times 10^{-5}}{0.0066}\right) = 0.85$$

Thus, 85% of all present CHH alleles are estimated to be direct descendants of the founder copy, implying $c = 0.12$ cM. Taking $\mu = 1 \times 10^{-6}$ gives $\alpha = 0.98$ and $c = 0.27$ cM. By contrast, traditional pedigree-based mapping was able to localize the CHH gene to only a 1.7 cM region.

LD mapping has been successfully applied to other disease genes segregating in the Finnish population (Peltonen et al. 1999), such as congenital nephrotic syndrome (Kestilä et al. 1994) and progressive myoclonus epilepsy (Lehesjoki et al. 1993). Kaplan et al. (1995) were able to apply LD mapping to the cystic fibrosis (CF) gene, using much more heterogeneous populations from Europe and elsewhere. The likely reason for success is that 70% of CF chromosomes worldwide appear to result from a single three-base deletion (Kerem et al. 1989), so that allelic heterogeneity is, at worst, a modest problem. Kaplan et al. (1995) found, however, that LD mapping was not very successful for Huntington's disease (multiple ancestral haplotypes) or Friedreich ataxia (high allelic heterogeneity) using European or North American populations.

Admixture Mapping

One generator of LD that has been exploited in mapping is **admixture**, the recent mixing of two populations that differ in allele frequencies. Suppose a sample of gametes comes from two distinct populations, where m is the fraction from population 1 and $(1-m)$ the fraction from population 2. Even if both populations are in LE, if there are allele frequency differences between them at loci A and B ($\delta_A, \delta_B \neq 0$), then the amount of LD between A and B in the sample is

$$D^{(0)} = m(1-m)\delta_A\delta_B \tag{17.20a}$$

Hence, the more even the mixing (m closer to 0.5) and the greater the allele frequency differences, the larger the initial LD. More generally, if D_i is the disequilibrium in population i, then (Chakraborty and Weiss 1988)

$$D^{(0)} = mD_1 + (1-m)D_2 + m(1-m)\delta_A\delta_B \tag{17.20b}$$

If t generations of random mating follow the initial admixture event, then (Equation 5.12b)

$$D^{(t)} = (1-c)^t D^{(0)} \tag{17.20c}$$

The historical interest in using such **admixture linkage disequilibrium** (**ALD**) was that when the admixture is recent (t in Equation 17.20c is small), regions that sufficiently differ in allele frequencies retain a LD signal over rather long regions (on the order of 10–30 cM). Hence, a genome-wide scan with a very modest number of markers is feasible. Conversely, in a random sample from most populations, LD exists only between *very tightly* linked markers (those separated by just a few kilobases, e.g., recombination frequencies of $<$ 0.1 cM), a feature that requires (at least) tens-of-thousands of SNPs to be exploited for association mapping (Chapter 20).

In the human genetics setting, Chakraborty and Weiss (1988) proposed that **admixture mapping** (**AM**) using the ALD generated during the contact between migrating populations could be used to map major genes that differ in allele frequencies between the mixing populations (an idea first suggested by Rife 1954). In particular, family-based tests (such as the TDT) that require sample-level LD are much more powerful when using admixed families, as these have much more widespread LD. Darvasi and Shifman (2005) noted that the admixture mapping strategy falls between the low-marker, poor resolution of linkage mapping and the high-marker, very fine resolution from association studies, with AM very much akin to using AIC lines for linkage-based QTL mapping (Chapter 18). Discussions of sample size and design consideration under AM are given by McKeigue (1997, 1998, 2005); Halder and Shriver (2003); Hoggart et al. (2004); Patterson et al. (2004); and Zhu (2004).

One of the first applications of a genome-wide AM scan was by Zhu et al. (2005), who examined hypertension in an admixture sample of African Americans, representing roughly 3/4 African and 1/4 European ancestry. They used 270 microsatellite markers to scan the genome, with a pooled sample of around 750 admixed cases (high blood pressure) and 570 admixed controls (normal pressure). Based on searching for an excess of African-ancestry markers in the cases versus the controls, they were able to suggest that regions on chromosomes 6 and 21 harbored QTLs that gave higher risk for hypertension. This study highlights one key limitation of AM, namely poor mapping resolution. As our ability to quickly and cheaply score millions of SNP markers evolved, the use of admixture mapping in humans quickly gave way to GWAS (Chapter 20), which offered much higher resolution under much less restrictive settings. However, hybrid zones between populations are common for most species in nature, and AM represents one potential strategy to exploit these settings.

CANDIDATE LOCI

In some cases, there may be sufficient physiological/biochemical information to suspect that certain known loci influence trait expression. In human genetics, it is common practice to directly test for population-level associations between trait value and specific alleles at such **candidate loci** (the **measured genotype** approach). For example, Boerwinkle and Sing (1987) showed that three common alleles at the human apolipoprotein E locus account for about 8% of the total variation in cholesterol levels. Interestingly, a particular allele of apolipoprotein E also appears to be a major determinant of Alzheimer's disease. The mean age of onset for homozygotes and heterozygotes for this allele is 68.4 ± 1.2 and 75.5 ± 1.0 years, respectively, while the mean age for individuals with no copies of this allele is 84.3 ± 1.3 (Corder et al. 1993). Other candidate loci for Alzheimer's disease are reviewed by Pericak-Vance and Haines (1995) and Selkoe (2001). As Example 17.17 shows, results using the candidate-locus approach can be rather unexpected.

How are candidate loci chosen? One obvious approach is to consider loci known from laboratory studies (such as knockout experiments) to have mutant alleles with major effects on the character of interest, as in natural populations such loci may also be segregating alleles with smaller effects (Chapter 15). Results from several QTL mapping experiments offer some support for this approach (Chapter 18). In maize, for example, Beavis et al. (1991), Edwards et al. (1992), Veldboom et al. (1994), and Berke and Rocheford (1995) found that many QTLs for height map near the locations of known height mutants. Similar findings have been obtained with bristle number in *Drosophila* by Mackay and Langley (1990), Lai et al. (1994), and Long et al. (1995). However, such finding are strongly influenced by conformation bias, as the true number of QTLs that do not show major knockout effects is unknown (although GWAS studies suggest it may be a considerable fraction, if not the majority; Chapters 20 and 21),

Example 17.17. Winkelman and Hodgetts (1992) examined the growth hormone (GH) gene as a candidate locus for body weight in selected lines of mice. Molecular analysis disclosed the presence of an allele, GH^h, present in all four lines selected for increased weight (being fixed in three of these). An alternative allele, GH^c, was fixed in all five control lines. One of the up-selected lines was crossed to two separate control lines to create two different F_2 populations. The GH^h allele had a significant, but unexpected, effect on body weight in both F_2 populations, as it *decreased* weight. For one population, the genotypes $GH^hGH^h : GH^hGH^c : GH^cGH^c$ had 42-day weights of 29.2 : 30.2 : 31.4, while in the other population these respective weights were 34.6 : 34.9 : 38.8. Thus, the GH^h allele was additive in one of the F_2 backgrounds, but dominant in the other. The GH^h allele behaved rather differently once the F_2 populations were subjected to selection, again increasing in frequency in up-selected lines, and decreasing in frequency in down-selected lines. Winkelman and Hodgetts suggested that the association between the GH^h allele and increased weight was a product of epistastic interactions in mice selected for high weight.

Example 17.18. Kozak et al. (2019) examined time to pupation under diapause-breaking photoperiod and temperature conditions (post-diapause development, PDD, time) in the European corn borer moth (*Ostrinia nubilalis*). Some eastern US populations have responded to shorter winters, evolving a PDD time that is roughly three weeks shorter than the late populations. A cross between an early and a late population found two significant QTLs, both located on the Z chromosome (recall that males in Lepidoptera are the homogametic sex; Chapter 29). The first QTL region (QTL1) was roughly 3.1Mb in size and contained 48 annotated genes, while the second region (QTL2) was 3.7Mb with 42 annotated genes. Further, the two regions showed a strong epistasic interaction for PDD. Both QTL regions contained a potential candidate gene: the circadian clock gene period (*per*) in QTL1 and pigment-dispersing factor receptor (*Pdfr*) in QTL2.

To further examine these potential candidates, the authors pooled sequence data from 5 populations, and measured marker allele frequency divergence among populations differing in PDD time. Across the entire genome, only six SNPs showed exceptional amounts of divergence (Bayes factors in excess of 20; Appendix 7), all of which fell inside the two QTL regions, with three of these within the *per* and *Pdfr* genes. Given the strong interaction between the two QTL regions, the authors expected the target sites of selection to likely be in high LD. In populations of this species, LD roughly spans 2 Mb or less, with sites between 2 and 10 Mb showing much lower amounts (99.9% have values of $r^2 \leq 0.56$). Among SNPs in annotated gene pairs located within or between the two QTL regions (distances of 2–7 Mb), they found 12 outlier pairs, the most extreme of which was between the *per* and *Pdfr* genes. Although roughly 5 Mb apart, they display an LD value of $r^2 = 0.75$. Taken together, these observations provides a strong case that *per* and *Pdfr* genes are major factors the evolution of PDD time in the early population.

The Transmission/Disequilibrium Test, TDT

When considering dichotomous (presence/absence) traits (such as disease status), the frequency of a particular candidate (or marker) allele in affected (or **case**) individuals is often compared with the frequency of this allele in unaffected (or **control**) individuals. The problem with such **association studies** is that a ***marker-trait association can arise simply as a consequence of population structure***, rather than as a consequence of linkage. Such **population stratification** occurs if the total sample consists of a number of divergent populations (e.g., different ethnic groups) which differ in both candidate-gene frequencies and incidences of the trait. Population structure can severely compromise tests of candidate gene associations, as the next example illustrates. As detailed in Chapter 20, a considerable amount of effort in any GWAS analysis is spent in attempts to control for any effects from population structure.

Example 17.19. Hanson et al. (1995) used segregation analysis (Chapter 16) to find evidence for a major gene for Type 2 diabetes mellitus segregating at high frequency in members of the Pima and Tohono O'odham nations of southern Arizona. In an attempt to map this gene, Knowler et al. (1988) examined how the simple presence/absence of a particular haplotype, Gm^+, was associated with diabetes. Their sample showed the following associations:

Gm^+	Total subjects	% with Diabetes
Present	293	8%
Absent	4,627	29%

The resulting χ^2 value (61.6, 1 df) shows a highly significant negative association between the Gm^+ haplotype and diabetes, making it very tempting to suggest that this haplotype marks a candidate diabetes locus (either directly or by close linkage).

However, the presence/absence of this haplotype is also a very sensitive indicator of admixture with the Caucasian population (who have a much lower diabete risk). The frequency of *Gm*$^+$ was around 67% in Caucasians as compared to $<$ 1% in full-heritage Pima and Tohono O'odham. When the authors restricted the analysis to such full-heritage adults (over age 35 to correct for age of onset), the association between haplotype and disease disappeared:

Gm$^+$	Total subjects	% with Diabetes
Present	17	59%
Absent	1,764	60%

Hence, the *Gm*$^+$ marker is a predictor of diabetes not because it is linked to genes influencing diabetes but rather because ***it serves as a predictor of whether individuals are from a specific subpopulation with a different disease risk***. *Gm*$^+$ individuals usually carry a significant fraction of genes of Caucasian extraction. Since a gene (or genes) increasing the risk of diabetes appears to be present at high frequency in individuals of full-heritage Pima/Tohono O'odham extraction, admixed individuals have a lower chance of carrying this gene (or genes).

The problem of population stratification can be overcome by employing tests that use *family data*, rather than *population data* from unrelated individuals, to provide the case and control samples (Penrose 1939; Woolf 1955; Clarke et al. 1956; Rubinstein et al. 1981; Falk and Rubinstein 1987; Terwilliger and Ott 1992; Spielman et al. 1993; Thomson 1995). One such family-based approach is to consider the transmission (or lack thereof) of a parental marker allele to an affected offspring. Focusing on *transmission within families* controls for associations generated entirely by population stratification and provides a direct test for linkage *provided* that a population-wide association between the marker and disease gene exists (Spielman et al. 1993; Ewens and Spielman 1995). This method can be applied to any family that has at least one affected offspring.

The **transmission/disequilibrium test**, or **TDT** (Spielman et al. 1993), compares the number of times a marker allele is transmitted (T) versus not-transmitted (NT) from a marker heterozygote parent to affected offspring. Under the hypothesis of no linkage, these values should be equal, and the test statistic becomes

$$\chi^2_{td} = \frac{(T - NT)^2}{(T + NT)} \tag{17.21}$$

which follows a χ^2 distribution with one degree of freedom. Equation 17.21 is also known as **McNemar's test of discordance** (Sokal and Rohlf 1995). There is a large literature on the TDT and its extensions (as a group, these methods are these are often called **family-based association tests**, **FBAT** or **family-based association analyses**, **FBAA**), with general overviews given by Schaid and Sommer (1994), Spielman and Ewens (1996), Schaid (1998), Rabinowitz and Laird (2000), Zhao (2000), Horvath et al. (2001), Schulze and McMahon (2002), Ewens et al. (2008), Laird and Lange (2006a, 2006b, 2009), Tiwari et al. (2008), Sneller et al. (2009), and Ott et al. (2011).

The signal from a TDT (a significant difference between T and NT) requires *both* linkage (an excess of parental gametes from a parent creating an association in their offspring between a marker allele and a DS allele) *and* population-level linkage disequilibrium (gametic phase largely agrees over families, allowing us to pool values over multiple families). Ewens and Spielman (1995) formalized this by developing expressions for $E[T - NT]$ under different admixture scenarios (parents of affected sibs are from a set of subpopulations, which are either distinct or have started to mix). All of their resulting expressions had the form of $2n \cdot (1 - 2c) \cdot \delta$, where n is the number of affected offspring from parents that are marker heterozygotes, c the marker-DS recombination frequency, and δ is a term that is only nonzero when population-level association (LD) occurs. For example, for affected offspring in the third generation after an admixture event,

$$E[T - NT] = 2n \frac{D}{p_{DS}} (1 - 2c) \tag{17.22}$$

where p_{DS} is the frequency of the disease allele, D the disequilibrium between the marker and DS allele, and c their recombination frequency. Hence, TDT has no power if $D = 0$ (no population-level disequilibrium between marker and disease alleles) or $c \simeq 1/2$ (no linkage between the marker and disease loci). As has been stressed by Ewens and Spielman (1995), the TDT is thus a *joint* test of *both* linkage *and* population-level disequilibrium, leading to its original description as a ***test for linkage in the* presence *of association*** (LD).

A number of papers have examined approximate and exact solutions for the power of a TDT and other FBAT approaches (Sham and Curtis 1995; Risch and Merikangas 1996; Camp 1997, 1999; Spielman and Ewens 1998; Whittaker and Lewis 1998; Knapp 1999b, 1999c; Tu and Whitemore 1999; McGinnis 2000; Chen and Deng 2001; Lange et al. 2002; Diao and Lin 2006; Fisher and Lewis 2008; Glaser and Holmans 2009). Under the null (*either* $c = 1/2$ or $D = 0$), Equation 17.21 is χ^2-distributed. When $c < 1/2$ *and* $D \neq 0$, Equation 17.21 now follows a noncentral χ^2 (Appendix 5), with noncentrality parameter (which determines the power of the test; Equaion A5.14a) being a function of c, the disequilibrium, and the nature of the disease (dominant, recessive, population frequency). Generally speaking, power is higher for rarer diseases and for recessive diseases. For example, Sham and Curtis (1995) using values for typical diseases found that the sample sizes (number of trios) required for 90% power (under the assumption that $c \simeq 0$) is around 100 for a rare recessive, 300 for a rare dominant, 400 for a common recessive, and over 1000 for a common dominant.

How are T and NT determined? Consider an M/m parent with three affected offspring. If two of those offspring received this parent's M allele, while the third received m, we score this as two transmitted M, one not-transmitted M. Conversely, if we are following marker m instead, this is scored as one transmitted m, two not-transmitted m. As the following example shows, each marker allele is examined separately under the TDT.

Example 17.20. Copeman et al. (1995) examined 21 microsatellite marker loci in 455 human families with Type 1 diabetes. One marker locus, *D2S152*, had three alleles, with one allele (denoted 228) showing a significant effect under the TDT. Parents heterozygous for this marker transmitted allele 228 to diabetic offspring 81 times, while transmitting alternative alleles only 45 times, giving

$$\chi^2 = \frac{(81-45)^2}{(81+45)} = 10.29$$

which has a corresponding *p* value of 0.001. As summarized below, the other two alleles (230 and 240) at this marker locus did not show a significant TD effect.

Allele	*T*	*NT*	χ^2	*p*
228	81	45	10.29	0.001
230	59	73	1.48	0.223
240	36	24	2.40	0.121

Hence, this marker appears to be linked to a QTL influencing Type 1 diabetes, with allele 228 in (coupling) linkage disequilibrium with an allele that increases the risk for this disease.

Spielman et al. (1993) noted that the TDT can give a false positive if the marker shows **segregation distortion**, wherein heterozygotes preferentially segregate one allele. This distortion can be controlled for by using a standard 2×2 contingency table χ^2 test, considering separately the transmission of a marker allele to affected and unaffected offspring,

	Transmitted	Not Transmitted
Affected Offspring	T_A	NT_A
Unaffected Offspring	T_U	NT_U

with resulting test statistic

$$\chi^2_{td} = \frac{(T_A - NT_A)^2}{(T_A + NT_A)} + \frac{(T_U - NT_U)^2}{(T_U + NT_U)} \tag{17.23}$$

controlling for any segregation distortion. A number of other linkage-based family methods for detecting marker-trait associations that do not require population-wide disequilibrium have been developed, and these are examined in Chapter 19. Chapter 20 examines the GWAS approach of exploiting population-wide disequilibrium between *extremely* close markers that can be detected using very dense marker maps.

The original TDT required **trio data**: genotypes from both parents and an affected offspring. In many cases, especially for diseases with a late age of onset, the genotypes of one (or both) parents may not be available. This issue has been addressed by a number of authors (Curtis and Sham 1995b; Curtis 1997; Laird et al. 1998; Spielman and Ewens 1998; Knapp 1999a, 1999c; Gordon et al. 2004), generally by using marker data from unaffected sibs in lieu of parental data, an approach Spielman and Ewens (1998) called **sib TDT** (or **S-TDT**). A related issue to missing genotypes are genotyping errors, which can impact the TDT. For example, Mitchell et al. (2003) noted that undetected genotyping errors create an apparent overtransmission of common alleles. Gordon et al. (2001, 2004) and Dudbridge (2008) introduced modifications of the TDT that are robust to random genotyping errors.

Other important extensions are to multiple marker loci, i.e., haplotypes (Bickeböller and Clerget-Darpoux 1995; Sham and Curtis 1995b; Clayton 1999; Clayton and Jones 1999; Dudbridge et al. 2000); and to application over pedigrees, as opposed to trio or sib data, leading to **pedigree disequilibrium tests**, **PDTs** (Abescasis et al. 2000; Martin et al. 2000, 2003; Gordon et al. 2004). Finally, while the classic TDT (and its above mentioned extension) applies to traits with present or absence count data (such as presence or absence of a disease), family-based segregation methods can be extended to quantitative (continuously valued) traits (Allison 1997; Allison et al. 1999; George et al. 1999; Sun et al. 2000; Zhu and Elston 2001; Malkin et al. 2002; Tiwari et al. 2005; Diao and Lin 2006). Such **quantitative transmission/disequilibrium tests**, or **QTDTs**, are reviewed by Ewens et al. (2008).

While the TDT, and its extensions, provide solid control over population structure, they require rather specialized samples, namely, family trios (or similar sets of relatives). As such, the total sample size in a TDT tends to be very small, resulting in modest to low power even when testing just a few potential candidates (unless their effect size is substantial). If one wishes to scan a large number of candidates (such as a set of SNPs covering a genome or genomic region), power is essentially non-existent, given the massive multiple-testing corrections required (Chapter 20). Hence, the TDT is best considered as a rather specialized, but gold-standard, test for replication of a signal from one (or a few) markers that are of very high interest (perhaps as a prelude to a rather expensive functional genomics or gene editing study). Chapter 21 examines other very-fine mapping methods when very dense SNP data is available.

Estimating Effects of Candidate Loci

While the estimation of genotype means for a candidate locus seems straightforward, there are several potential sources of bias. While it is possible in some settings to distinguish between the direct effects of a candidate locus and its indirect effects due to association with other linked QTLs (e.g., Bovenhuis and Weller 1994), as noted above the presence of linkage disequilibrium greatly confounds the interpretation of candidate-locus means (Chapter 21 examines methods for very fine mapping of actual causal sites using dense SNPs). We now consider more subtle sources of bias.

Let z_{ij} denote the phenotypic value of the jth individual with candidate-locus genotype i $(1 \leq i \leq n_g)$. If μ_i is the mean character value for genotype i, then the simplest model is

$$z_{ij} = \mu_i + \epsilon_{ij} \tag{17.24}$$

If individuals are sampled at random from a large homogeneous population, the residuals are uncorrelated and the μ_i values can be estimated simply by using $\overline{z}_i$. However, when relatives are present in the sample, residuals can be correlated, and one must take the residual covariance matrix into consideration to obtain an unbiased estimator (Example 10.12). The residual error can be decomposed into a genetic component (G) due to segregation at loci other than the target QTL plus the general (E) and specific (e) environmental effects,

$$\epsilon_{ij} = G_{ij} + E_{ij} + e_{ij} \tag{17.25a}$$

Indexing two distinct individuals by j and k,

$$\sigma(\epsilon_j, \epsilon_k) = \sigma(G_j, G_k) + \sigma(E_j, E_k) \tag{17.25b}$$

as special environmental effects are not shared by different individuals (Chapter 6), and G and E are assumed to be uncorrelated. If the relationships among sampled individuals are known, the elements of the covariance matrix can be computed from Equation 17.25b, using the methods of Chapter 7. For example, if i and j are full-sibs, $\sigma(G_j, G_k) = (\sigma^2_A/2) + (\sigma^2_D/4)$, where the variances refer to the contribution of background polygenes. Similarly, full sibs may also share a common material environmental effect (m), so that $\sigma(E_j, E_k) = \sigma^2_m$. Estimates of the variance components associated with G and E can be obtained by a variety of methods introduced in Chapters 22–33. Using these to estimate the residual covariance matrix, genotypic means can be estimated by a GLS regression (Equation 10.25e).

An example of the importance of correcting for common relatives is provided by Bentsen and Klemetsdal (1991), who examined the association between egg productivity and six different MHC haplotypes in chickens. The relative rankings of 5 of the 6 haplotype effects differed when comparing uncorrected estimates with those obtained by correcting for shared ancestry. More generally, when additional fixed factors (such as sex- or age-specific effects) are included, mixed-model methods can be used (Boerwinkle et al. 1986; Cowan et al. 1990; Hoeschele and Meinert 1990; Bentsen and Klemetsdal 1991; Blangero et al. 1992; Kennedy et al. 1992). Chapter 31 examines this in some detail.

Many estimates of the amount of variability attributable to a candidate locus are upwardly biased, even for a random population sample. If q_i is the frequency of candidate-locus genotype i, then the variance contributed by this locus is

$$\sigma^2_L = \sum_{i=1}^{n_g} q_i\,(\,\mu_i - \mu\,)^2 \qquad \text{where} \qquad \mu = \sum_{i=1}^{n_g} q_i\,\mu_i \tag{17.26}$$

Boerwinkle and Sing (1986) showed that the obvious estimator

$$s^2 = \sum_{i=1}^{n_g} \widehat{q}_i\,(\,\overline{z}_i - \overline{z}\,)^2 \tag{17.27a}$$

using the estimates $\widehat{q}_i$, $\overline{z}_i$, and $\overline{z}$ in place of their true values is biased because it includes the sampling error of these three estimates. To account for the sampling variances in $\overline{z}_i$ and $\overline{z}$, Boerwinkle and Sing suggest the improved estimator (for a collection of unrelated individuals),

$$s^2_L = \sum_{i=1}^{n_g} \widehat{q}_i\,(\,\overline{z}_i - \overline{z}\,)^2 - \left(\frac{n_g - 1}{n}\right)\text{Var}(e) \tag{17.27b}$$

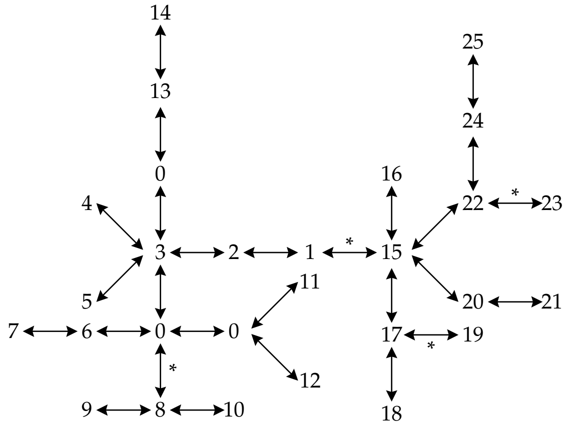


Figure 17.11 Inferred cladogram of haplotypes for a region surrounding the *Drosophila melanogaster* ADH gene. Numbers refer to different haplotypes, while "0" refers to intermediate haplotypes not present in the sample but required to interconnect the existing haplotypes. Distance between two haplotypes on the cladogram provides a measure of relatedness, so that haplotypes 8 and 9 are more closely related than 9 and 6 (one step vs. three steps). Using the hierarchical structure implied by this cladogram, a nested ANOVA was performed, with ANOVA levels set by the number of cladogram steps that haplotypes differ by. Four cladogram regions, denoted with asterisks, were found to be associated with significant phenotypic effects. (After Templeton et al. 1987.)

where Var(e) is the estimate of the residual variance (within genotypes) and we assume n measured individuals per genotype. While an improvement over Equation 17.27a, this estimator still ignores the sampling variance introduced by using $\widehat{q}_i$ and hence still tends to overestimate the variance accounted for by the candidate locus. Resampling approaches for constructing confidence intervals of both σ^2_L and its additive and dominance components have been suggested by Kaprio et al. (1991) and Zerba et al. (1996).

A general solution to this reoccurring issue of estimating the sample variance of a composite parameter (σ^2_L) that itself is a function of other parameters that have sampling errors (q_i, μ_i) is to use a full **Bayesian analysis** (Appendices 7 and 8). Under such an analysis, one obtains the posterior distribution of the vector of parameters $(q_1, \cdots q_{n_g}, \mu_1, \cdots, \mu_{n_g})$ by starting with some prior belief about their values (which is often a flat prior), and then use the observed data values to update this distribution to generate a posterior. As Appendix 8 details, this is usually done by using Markov Chain Monte Carlo methods (MCMC) to generate draws from this posterior distribution. Under such an analysis, a drawn from the MCMC sampler returns a vector of values for the q_i and μ_i which are then substituted into Equation 17.26 to generate a corresponding estimate σ^2_L. One repeats this process several thousand times to generate an empirical distribution of σ^2_L estimates that fully account for the sampling value in both the q_i and μ_i.

Templeton and Sing's Method: Using the Historical Information in Haplotypes

While one can use a single marker to examine whether genetic variation at a candidate locus is associated with character variation, studies typically involve several closely linked markers, often including several polymorphic sites within the gene itself. How should one best extract information from this set of markers? Obviously, it is more powerful to consider **haplotypes** (the multiple-locus genotypes associated with the region of interest) than single markers. One drawback with this approach is that there can be an enormous number of haplotypes, resulting in small sample sizes for each haplotype and a reduction in the power of tests for haplotype-trait associations. What is needed is a logical way to combine information from different haplotypes. In an interesting series of papers, Templeton and Sing (Templeton et al. 1987, 1988, 1992; Templeton 1995) suggest that this can be done by incorporating information on the inferred evolutionary relationships of the sampled haplotypes. We simply outline their basic idea here.

The sequence information in haplotypes can be used to construct a **cladogram** estimating how the different haplotypes are evolutionarily related to each other. With such a cladogram in hand, one can then consider how character value is a function of nested sets of cladogram members, for example by using a nested ANOVA (Chapter 23). The motivation behind this approach is that history is the main cause of linkage disequilibrium, with associations between loci decaying with time. Hence, the more closely related a set of haplotypes, the more disequilibrium they should display, and the more likely they are to share QTL alleles. Figure 17.11 shows an example of this approach, detecting associations between alcohol dehydrogenase (ADH) activity and haplotypes in a 13 kb region surrounding the ADH gene in *Drosophila*. This approach was used by Klein et al. (2005) in one of the first GWAS studies (Chapter 20) as part of their evidence that a polymorphism in human Complement Factor H was associated with age-related macular degeneration.