11

Analysis of Line Crosses

Version 12 Nov 2020

Distinct populations, such as the isolated demes that comprise some natural populations or local land races and breeds of domesticated species, often exhibit remarkable phenotypic divergence. Such differences are sometimes a simple consequence of environmental influences on phenotypic expression, but genetic differences may arise as a result of local adaptation and random genetic drift. The genetic basis of interdemic differentiation is of interest for several reasons. With nonadditive gene action, the mean phenotypes of progeny of interdemic crosses (F_1 hybrids from a line cross) will not be intermediate to those of their parents. One manifestation of this phenomenon is **heterosis**, the common observation that hybrids in many domesticated species often have enhanced performance or trait values relative to their parental lines (Chapter 13). Conversely, a problem of some significance in conservation biology is the opposite observation: **outbreeding depression**, wherein hybrids have *reduced* fitness relative to their parental populations (Chapter 13). Thus, depending on circumstances, line crossing can either very desirable (heterosis) or to be avoided as much as possible (outbreeding depression). For populations that do not normally have an opportunity to interbreed in nature, the nature of their potential hybrids may evolve passively as an indirect consequence of local adaptation and drift. However, when opportunities for interdemic exchange are common, natural selection may favor specific mating system properties, including dispersal strategies or reproductive isolation, which enhance or discourage outcrossing (a key to deciphering the mechanisms of speciation).

The mechanisms of genetic differentiation also have important practical implications beyond prediction of hybrid performance. For example, in artificial selection programs, the mean phenotypes of selected lines often evolve well beyond the range of variation seen in the base population prior to selection (Chapter 1; WL Chapters 25 and 26). The extent to which such changes are caused by a large number of genes of relatively small effects, as opposed to a few major segregating factors, is an important determinant of whether a search for informative molecular markers is likely to be successful (Chapters 17–20). In addition, the degree to which selection advances made in different lines can be successfully integrated into a single crossbred line is a function of the ways in which genes from the isolated lines interact. Entirely additive gains are easy to stack across lines, while incorporating gains from nonadditive interactions is more problematic.

In this chapter, we show how the judicious choice of line crosses can be used to reveal the relative contributions of additive, dominance, and epistatic effects to population differentiation. A statistical test of the adequacy of alternative genetic models will be presented, and its application to a variety of data sets will be used to show that nonadditive gene action is commonly associated with population differentiation. Several methods for estimating the minimum number of loci responsible for population differentiation will then be discussed. Their application firmly supports the conviction that most characters of interest to evolutionary biologists and breeders are influenced by multiple loci. This, however, does not rule out the possibility that a small number of loci are responsible for the majority of the differentiation between species and/or lines within species. Chapter 13 examines both heterosis and outbreeding depression, while Chapter 12 examines the complementary concern of inbreeding depression, the decay in performance upon inbreeding.

EXPECTATIONS FOR LINE-CROSS MEANS

We start with two parental populations (P_1 and P_2), each with loci assumed to be in Hardy-

Weinberg and gametic phase equilibrium (note that a fully inbred line is in Hardy-Weinberg, as all loci are fixed are a single allele). For the time being, we also assume that the loci differentiating the two populations are unlinked. An F_1 population is obtained by crossing the P_1 and P_2 lines, and subsequent random mating of F_1 individuals results in an F_2 generation. Because the F_2 population will be in Hardy-Weinberg and gametic phase equilibrium for unlinked loci, both for genes derived within and between populations, it is logical to treat it as a point of reference for the definition of genetic effects.

In previous chapters, we have presented definitions of additive, dominance, and epistatic effects for specific genes and gene combinations within randomly mating populations. The statistical machinery underlying line-cross analysis has parallels with this approach. The questions here, however, are whether there is a net difference between the additive effects of the genes in the P_1 and P_2 populations, whether the genes in P_1 tend on average to be dominant over those in P_2, and whether there are net directional epistatic interactions between P_1 and P_2 genes. Before explicitly defining these **composite effects**, it will be useful to have some indices to describe the **gene content** and the **degree of hybridity** of line-cross derivatives.

For any pair of lines and their derivatives, the genotypes at a locus can be partitioned into three classes: (1) both alleles are from a random sample of P_1 genes, (2) both alleles are from a random sample of P_2 genes, and (3) one allele is from a random sample of P_1 genes, while the other is from the P_2 pool of genes. Let S be the fraction of P_1 genes in a line, and H be the probability that a member of the line has one P_1 and one P_2 gene at a locus. These two indices uniquely specify the expected frequencies of the three classes of genotypes at any autosomal locus:

$$\begin{aligned} S - \frac{H}{2} &= \text{frequency of individuals containing only } P_1 \text{ alleles} \\ H &= \text{frequency of individuals containing one } P_1 \text{ and one } P_2 \text{ allele} \\ 1 - S - \frac{H}{2} &= \text{frequency of individuals containing only } P_2 \text{ alleles} \end{aligned} \qquad (11.1a)$$

Composite effects are formally defined in a least-squares framework, similar to that used for effects at single loci (Chapter 4). Consider first the **composite additive effect**, defined as the difference of additive effects of P_1 versus P_2 alleles summed over all loci. A simple expression for this can be obtained by recalling Equation 4.9—for a diallelic locus, the additive effects of the B_1 and B_2 alleles can be written as $-p_2\alpha$ and $p_1\alpha$, where α is the average effect of allelic substitution, and p_1 and p_2 are the frequencies of the B_1 and B_2 alleles. These expressions apply to a randomly mating population, precisely the situation in the F_2 generation. Noting that the F_2 consists of 50% P_1 and 50% P_2 genes, the frequencies of all contrasting alleles (of P_1 versus P_2 origin) are 0.5. Thus, the composite additive effects of P_1 and P_2 genes in the F_2 reference population are equal in absolute value but opposite in sign, and we denote them respectively as $+\alpha^c/2$ and $-\alpha^c/2$, where the superscript c denotes composite (as opposed to single-gene) effects. The total composite additive effect in the F_2 generation, $[(\alpha^c - \alpha^c)/2]$, is then equal to zero, which is what is desired for a reference population. The total composite additive effect in the F_1 generation is also equal to zero, because every locus contains one P_1 and one P_2 allele. In the P_1 population, however, there are only P_1 alleles, each of which contributes $\alpha^c/2$, giving the total composite additive effect as $(\alpha^c + \alpha^c)/2 = \alpha^c$. In contrast, as each P_2 allele contributes $-\alpha^c/2$, the total composite effect in the P_2 population is $-\alpha^c$. More generally, the contribution of composite additive effects to the mean genotypic value of any line-cross derivative is

$$\left(S - \frac{H}{2}\right)(+\alpha^c) + (H)(0) + \left(1 - S - \frac{H}{2}\right)(-\alpha^c) = (2S-1)\alpha^c = \theta_S\alpha^c \qquad (11.1b)$$

where $\theta_S = 2S - 1$ denotes the **source index.** The three terms on the left represent, respectively, the contributions from P_1 homozygotes, P_1P_2 heterozygotes, and P_2 homozygotes.

The index θ_S contrasts the expected number of P_1 alleles at a locus in a particular line $(2S)$ with that in the F_2 reference population (1).

The **composite dominance effect** is obtained in a similar manner, again treating the F_2 population as a reference. Returning to Table 4.2, it can be shown that when the frequencies of the two alleles are equal to 0.5, the dominance deviations from the regression on gene content are $-d/2$ for both homozygotes and $d/2$ for the heterozygote. An analogous situation exists for the F_2 of a line cross, which consists of 25% P_1P_1, 50% P_1P_2, and 25% P_2P_2 individuals at each locus. Thus, we denote the composite dominance effects associated with crossbred (P_1P_2) and purebred $(P_1P_1,\ P_2P_2)$ loci as $+\delta^c$ and $-\delta^c$, respectively. More generally, the contribution of composite dominance effects to the genotypic mean of a particular line is

$$\left(S-\frac{H}{2}\right)(-\delta^c)+(H)(+\delta^c)+\left(1-S-\frac{H}{2}\right)(-\delta^c)=(2H-1)\delta^c=\theta_H\delta^c \qquad (11.1c)$$

where $\theta_H = 2H - 1$ is the **hybridity index.** Thus, $\theta_H\delta^c$ yields a value of δ^c for the F_1 population, which consists entirely of hybrids, $-\delta^c$ for the parental lines, and zero for the F_2, maintaining the property that the average composite effects are defined to be zero in the reference population.

There are three important points to note about this approach to interpreting line-cross means. First, the composite effects are denoted as such because they *summarize the total effects over all loci*. Because some of the effects of individual alleles in population P_1 may be positive and others negative, there is a possibility of considerable cancellation of locus-specific effects. A comparison of the variances within different line-cross derivatives can shed some light on this problem (Mather and Jinks 1982), but we will not take this up here.

Second, provided there is no mating with close relatives, the definition of composite effects does not require that the parental populations be pure (completely homozygous) lines. For any line-cross derivative, the subset of individuals with two P_1 alleles at a particular locus will have the same expected Hardy-Weinberg genotype distribution of P_1 genotypes as the P_1 generation. The same argument applies to loci with two P_2 alleles. Thus, in the absence of inbreeding (which alters the genotypic frequencies within classes; see Chapter 12), differences between the means of various line-cross derivatives cannot be due to a shift in the genotype frequencies within the groups P_1P_1, P_1P_2, and P_2P_2. It can only be caused by a shift in the *relative abundances* of these three groups.

Third, the source and hybridity indices are all that are needed to define the contributions of composite epistatic effects to line means. To obtain the general expression, we let $(\alpha_n^c\delta_m^c)$ denote the composite effect involving the interaction of n additive and m dominance effects. (Note that this notation does not imply that $(\alpha_n^c\delta_m^c)=\alpha_n^c\cdot\delta_m^c$.) The general expression for the mean genotypic value of a line is then

$$\mu=\mu_0+\theta_S\alpha_1^c+\theta_H\delta_1^c+\theta_S^2\alpha_2^c+\theta_S\theta_H(\alpha_1^c\delta_1^c)+\theta_H^2\delta_2^c+\cdots \qquad (11.1d)$$

where μ_o is the F_2 mean. Note that the coefficients for the composite effect $\alpha_n^c\delta_m^c$ is of the form of $\theta_S^n\,\theta_H^m$. Hill (1982a) provides a formal derivation of Equation 11.1d, and alternative modes of presentation appear in Cockerham (1980), Lynch (1991), and Schnell and Cockerham (1992). Example 11.1 provides a derivation of the result for the additive × additive component of epistasis. The compositions of the expected line means for common types of crossbreds are given in Table 11.1.

The preceding model applies in the presence of linkage as long as there is no epistasis. Linked genes tend to be inherited as a unit, although they are gradually rendered independent by crossing-over. This process has the effect of altering the likelihood of specific epistatic interactions through progressive rounds of recombination (see Equations 11.2a–11.2c). Provided the parental lines are in gametic phase equilibrium, the expressions for the P_1, P_2, and F_1 lines still hold, because there is no opportunity for recombination between chromosomes of different parental lines, but those for all subsequent line-cross derivatives

Table 11.1 Expected mean phenotypes for various line-cross derivatives in terms of composite additive, dominance, and two-locus epistatic effects, taking the F_2 mean (μ_0) as a point of reference. These expressions assume freely recombining loci. For situations involving self-fertilization, it is further assumed that the parental lines are completely homozygous.

Line	θ_S	θ_H	Expected Mean Phenotype
P_1	1	–1	$\mu_0 + \alpha_1^c - \delta_1^c + \alpha_2^c - \alpha_1^c\delta_1^c + \delta_2^c + \cdots$
P_2	–1	–1	$\mu_0 - \alpha_1^c - \delta_1^c + \alpha_2^c + \alpha_1^c\delta_1^c + \delta_2^c + \cdots$
F_1	0	1	$\mu_0 + \delta_1^c + \delta_2^c + \cdots$
F_2	0	0	μ_0
$B_1 = (P_1 \times F_1)$	$\frac{1}{2}$	0	$\mu_0 + \frac{1}{2}\alpha_1^c + \frac{1}{4}\alpha_2^c + \cdots$
$B_2 = (P_2 \times F_1)$	$-\frac{1}{2}$	0	$\mu_0 - \frac{1}{2}\alpha_1^c + \frac{1}{4}\alpha_2^c + \cdots$
$F_2 \times P_1$	$\frac{1}{2}$	0	$\mu_0 + \frac{1}{2}\alpha_1^c + \frac{1}{4}\alpha_2^c + \cdots$
$F_2 \times P_2$	$-\frac{1}{2}$	0	$\mu_0 - \frac{1}{2}\alpha_1^c + \frac{1}{4}\alpha_2^c + \cdots$
$F_2 \times F_1$	0	0	μ_0
$B_1 \times F_1$	$\frac{1}{4}$	0	$\mu_0 + \frac{1}{4}\alpha_1^c + \frac{1}{16}\alpha_2^c + \cdots$
$B_2 \times F_1$	$-\frac{1}{4}$	0	$\mu_0 - \frac{1}{4}\alpha_1^c + \frac{1}{16}\alpha_2^c + \cdots$
Second backcrosses			
$B_1 \times P_1$	$\frac{3}{4}$	$-\frac{1}{2}$	$\mu_0 + \frac{3}{4}\alpha_1^c - \frac{1}{2}\delta_1^c + \frac{9}{16}\alpha_2^c - \frac{3}{8}\alpha_1^c\delta_1^c + \frac{1}{4}\delta_2^c + \cdots$
$B_1 \times P_2$	$-\frac{1}{4}$	$\frac{1}{2}$	$\mu_0 - \frac{1}{4}\alpha_1^c + \frac{1}{2}\delta_1^c + \frac{1}{16}\alpha_2^c - \frac{1}{8}\alpha_1^c\delta_1^c + \frac{1}{4}\delta_2^c + \cdots$
$B_2 \times P_1$	$\frac{1}{4}$	$\frac{1}{2}$	$\mu_0 + \frac{1}{4}\alpha_1^c + \frac{1}{2}\delta_1^c + \frac{1}{16}\alpha_2^c + \frac{1}{8}\alpha_1^c\delta_1^c + \frac{1}{4}\delta_2^c + \cdots$
$B_2 \times P_2$	$-\frac{3}{4}$	$-\frac{1}{2}$	$\mu_0 - \frac{3}{4}\alpha_1^c - \frac{1}{2}\delta_1^c + \frac{9}{16}\alpha_2^c + \frac{3}{8}\alpha_1^c\delta_1^c + \frac{1}{4}\delta_2^c + \cdots$
Selfed backcrosses			
B_{1s}	$\frac{1}{2}$	$-\frac{1}{4}$	$\mu_0 + \frac{1}{2}\alpha_1^c - \frac{1}{4}\delta_1^c + \frac{1}{4}\alpha_2^c - \frac{1}{8}\alpha_1^c\delta_1^c + \frac{1}{16}\delta_2^c + \cdots$
B_{2s}	$-\frac{1}{2}$	$-\frac{1}{4}$	$\mu_0 - \frac{1}{2}\alpha_1^c - \frac{1}{4}\delta_1^c + \frac{1}{4}\alpha_2^c + \frac{1}{8}\alpha_1^c\delta_1^c + \frac{1}{16}\delta_2^c + \cdots$
Continued selfing from the F_2			
F_3	0	$-\frac{1}{2}$	$\mu_0 - \frac{1}{2}\delta_1^c + \frac{1}{4}\delta_2^c + \cdots$
F_4	0	$-\frac{3}{4}$	$\mu_0 - \frac{3}{4}\delta_1^c + \frac{9}{16}\delta_2^c + \cdots$

will be biased to an extent depending on the map structure of the constituent loci. For a pair of loci with recombination fraction c, the following modifications need to be applied to the expressions for the F_2 and backcross means,

$$\mu(F_2) = \mu_0 + \left(\frac{1-2c}{2}\right)\alpha_2^c + (1-2c)^2\delta_2^c + \cdots \tag{11.2a}$$

$$\mu(B_1) = \mu_0 + \frac{\alpha_1^c}{2} + \left(\frac{1-c}{2}\right)\alpha_2^c + \left(\frac{2c-1}{2}\right)\alpha_1^c\delta_1^c + (1-2c)\delta_2^c + \cdots \tag{11.2b}$$

$$\mu(B_2) = \mu_0 - \frac{\alpha_1^c}{2} + \left(\frac{1-c}{2}\right)\alpha_2^c - \left(\frac{2c-1}{2}\right)\alpha_1^c\delta_1^c + (1-2c)\delta_2^c + \cdots \tag{11.2c}$$

Table 11.2 Genetic map structure and the average recombination frequency $(\bar{c})$ between random pairs of loci throughout the genome. N is the haploid number of chromosomes per genome, and L is the total map length (in Morgans) in females. For *Drosophila,* we ignore a tiny dot chromosome for which little mapping data are available. In the mosquito *Aedes* and in humans, respectively, recombination in males is approximately 50% and 65% as frequent as that in females; and there is no recombination within chromosomes in male *Drosophila.* For these three taxa, the reported values of $\bar{c}$ are averages for the two sexes. All results are approximations, as genomic maps are continuously being refined with the addition of molecular markers (Chapter 17).

Species	N	L	Lengths of Individual Chromosomes, L_i	$\bar{c}$
Drosophila melanogaster	3	2.77	0.66, 1.03, 1.08	0.365
Drosophila pseudoobscura	4	4.46	0.68, 0.69, 1.01, 2.08	0.386
Aedes aegypti	3	2.28	0.62, 0.80, 0.86	0.380
Caenorhabditis elegans	5	1.61	0.27, 0.31, 0.33, 0.34, 0.36	0.418
Arabidopsis thaliana	5	5.24	0.63, 0.83, 0.98, 1.36, 1.44	0.443
Hordeum vulgare (barley)	7	9.49	0.70, 1.12, 1.15, 1.26, 1.27, 1.59, 2.40	0.465
Neurospora crassa	7	10.02	1.07, 1.13, 1.18, 1.33, 1.49, 1.52, 2.30	0.466
Zea mays (maize)	10	12.10	0.42, 0.78, 0.95, 1.07, 1.12, 1.37, 1.41, 1.55, 1.55, 1.67, 1.76	0.474
Phaseolus vulgaris (bean)	11	8.92	1.05, 1.04, 0.95, 0.92, 0.86, 0.78, 0.74, 0.71, 0.71, 0.60, 0.56	0.471
Pinus pinaster (maritime pine)	12	18.56	1.90, 1.75, 1.69, 1.66, 1.63, 1.61, 1.57, 1.54, 1.54, 1.41, 1.36, 0.90	0.481
Lycopersicon esculentum (tomato)	12	14.91	0.90, 0.92, 0.98, 1.01, 1.04, 1.04, 1.23, 1.34, 1.42, 1.63, 2.11	0.479
Mus musculus (mouse)	20	14.25	0.36, 0.36, 0.49, 0.56, 0.57, 0.57, 0.68, 0.70, 0.71, 0.73, 0.74, 0.78, 0.78, 0.80, 0.81, 0.84, 0.87, 0.89, 1.00, 1.01	0.483
Homo sapiens	23	40.00	0.69, 0.76, 1.12, 1.12, 1.22, 1.24, 1.25, 1.26, 1.38, 1.39, 1.40, 1.47, 1.62, 1.67, 1.67, 1.67, 1.74, 1.75, 1.75, 1.77, 1.92, 2.21, 2.49	0.490
Danio rerio (zebrafish)	25	28.06	0.59, 0.80, 0.84, 0.85, 0.86, 0.87, 0.95, 1.00, 1.02, 1.05, 1.05, 1.09, 1.13, 1.13, 1.14, 1.16, 1.22, 1.22, 1.33, 1.34, 1.39, 1.39, 1.45, 1.53, 1.66	0.489

Source: All data are from O'Brien (1990), except that for *Phaseolus* (Vallejos et al. 1992), *Pinus* (Plomion et al. 1995), and *Danio* (supplied by J. Postlethwait).

Example 11.1 presents the derivation of Equations 11.2a–11.2c. The expressions for more advanced crosses (e.g., $B_1 \times F_1$) are complicated because one must account for additional generations of recombination.

Assuming that the epistatic effects between loci are independent of c, these expressions also apply to the total composite effects of all loci when $\bar{c}$, the mean recombination frequency between all pairs of loci, is substituted for c. Because we generally do not know c for any pair of loci, let alone for all of the loci underlying a quantitative trait, the best we can provide is a heuristic guide to the potential significance of linkage. Assuming that the genes are uniformly distributed across all chromosomes, then

$$\bar{c} = 0.5 - \frac{2L - N + \sum_{i=1}^{N} e^{-2L_i}}{4L^2} \tag{11.3}$$

where N is the number of chromosomes in a haploid set, L_i is the genetic map length of the ith chromosome (in Morgans), and $L = \sum L_i$ is the total map length (Zeng et al. 1990). This expression is based on Haldane's mapping function, $c_{ij} = 0.5(1 - e^{-2L_{ij}})$, which relates the recombination frequency to the genetic map length between two linked loci, i and j (Chapter 17). Estimates of $\bar{c}$ are given in Table 11.2 for several species for which the genetic maps are reasonably well resolved. For species with a very small number of chromosomes and/or restricted recombination in males, as in *Drosophila* and the mosquito *Aedes*, $\bar{c}$ can be somewhat less than 0.40, but when N exceeds six or so, it tends to be greater than 0.45. For most mammals, N is typically on the order of 15 or more, so most pairs of genes are on different chromosomes, and $\bar{c}$ is very close to 0.5. With $\bar{c} > 0.45$, Equations 11.2a-11.2c are quite close to the expressions in Table 11.1. Thus, unless the genes underlying quantitative traits tend to be aggregated on chromosomes, linkage is unlikely to cause much bias in the interpretation of line-cross means, except perhaps in the case of species such as *Drosophila.*

Example 11.1. All of the composite effects described above are defined in a least-squares sense, and the nice symmetry whereby all effects have the same absolute value but differ in sign is a consequence of all contrasting pairs of alleles having frequency 0.5 in the F$_2$ generation. We now provide a formal derivation of the additive × additive composite effects in the context of a reference population that is both in Hardy-Weinberg and gametic phase equilibrium. We denote the four gamete types as x_{11}, x_{12}, x_{21}, and x_{22}, where the subscripts refer to the parental sources of alleles (lines 1 or 2)at the first and second locus. All four gamete types have frequencies equal to 0.25. Let the additive × additive effects associated with these gametes be α_{11}, α_{12}, α_{21}, and α_{22}. The effect α_{ij} is defined to be the average residual effect associated with a gamete containing a P$_i$ -derived allele at the first locus and a P$_j$ -derived allele at the second locus, after the additive effects of the two genes have been accounted for (see Equation 5.4a). Under a least-squares framework, the mean residual error is defined to be zero (Chapter 3), which implies

$$\frac{1}{4}\left(\alpha_{11} + \alpha_{12} + \alpha_{21} + \alpha_{22}\right) = 0$$

Furthermore, the mean squared error is minimized. Noting that the previous expression implies that $\alpha_{22} = -\alpha_{11} - \alpha_{12} - \alpha_{21}$, the function to be minimized is

$$M = \alpha_{11}^2 + \alpha_{12}^2 + \alpha_{21}^2 + (-\alpha_{11} - \alpha_{12} - \alpha_{21})^2$$

Taking the partial derivative with respect to α_{11} and setting it equal to zero, we obtain

$$2\alpha_{11} + \alpha_{12} + \alpha_{21} = 0$$

Subtracting $(\alpha_{11} + \alpha_{12} + \alpha_{21} + \alpha_{22})$ from this expression, we find that the epistatic effects associated with each of the parental chromosome types are equal, i.e., $\alpha_{11} = \alpha_{22}$. By similar means, it can be shown that $\alpha_{12} = \alpha_{21}$, which when applied to the constraint that $(\alpha_{11} + \alpha_{12} + \alpha_{21} + \alpha_{22}) = 0$ implies that

$$\alpha_{11} = \alpha_{22} = -\alpha_{12} = -\alpha_{21}$$

Thus, the additive × additive effects associated with both recombinant chromosome types are equal and opposite in sign to those of the parental chromosomes. Because, in a diploid, there are four combinations of genes at two loci (two within and two between the uniting gametes), we define the effects as

$$\alpha_{11} = \alpha_{22} = +\alpha_2^c/4$$
$$\alpha_{12} = \alpha_{21} = -\alpha_2^c/4$$

In both the P_1 and P_2 populations, all additive × additive interactions within and between chromosomes are of parental type (α_{ii}), and the composite effect of such interactions is $4(\alpha_2^c/4) = \alpha_2^c$. In the F_1 generation, the two interactions within chromosomes are of parental type, but the pairs of nonalleles between chromosomes are of different parental type, so the composite effect is $2(\alpha_2^c/4) + 2(-\alpha_2^c/4) = 0$.

Now consider the situation in the F_2 generation. Under free recombination, all four gametes are equally frequent, and with random mating, the average additive × additive epistatic effect within and between uniting gametes is equal to zero, as noted above. Suppose, however, that the two loci are linked, so that a fraction c of gametes are recombinant, and a fraction $(1 - c)$ are nonrecombinant. The gamete frequencies are then $p(x_{12}) = p(x_{21}) = c/2$, and $p(x_{11}) = p(x_{22}) = (1 - c)/2$. Assuming random mating, the paternal source of an individual's allele at one locus is independent of the maternal source of an allele at a second locus. Thus, the average composite effects associated with additive × additive interactions between uniting gametes is equal to zero. However, because the parental sources of genes *within* gametes will not have been completely randomized, the composite additive × additive effect within gametes is not zero, but

$$2[c(-\alpha_2^c/4) + (1 - c)(\alpha_2^c/4)] = (1 - 2c)\alpha_2^c/2$$

giving the term in Equation 11.2a. These results can be extended to the backcross generations. xonsider, for example, the situation when F_1 individuals are crossed to members of parental line P_1, creating the B_1 backcross generation. The F_1 parent can contribute each of the four possible gametes, while the P_1 parent contributes only x_{11} gametes. With probability $(1 - c)/2$, the offspring genotype is x_{11}/x_{11}, and each of the four possible two-locus interactions involves two P_1 alleles; each such interaction contributes $\alpha_2^c/4$, giving $(1 - c)\alpha_2^c/2$. Likewise, with probability $c/2$ the genotype is x_{12}/x_{11}. Here, there are two P_1/P_1 interactions $(2\alpha_2^c/4)$ and two P_1/P_2 interactions $(-2\alpha_2^c/4)$, yielding a total contribution for this genotype of $(c/2)(2\alpha_2^c/4 - 2\alpha_2^c/4) = 0$. In a similar fashion, expectations for the other two genotypes are found to be equal to zero, giving the total contribution in the B_1 backcross as $(1 - c)\alpha_2^c/2$.

ESTIMATION OF COMPOSITE EFFECTS

With the preceding model, the expected line means are linear functions of the composite effects. Thus, straightforward procedures can be used to estimate these effects from the observed line means. Generally, when the estimates of k parameters are desired, k types of lines can be identified that allow a solution using simultaneous equations. For example, if the epistatic effects are ignored, the expectations for the P_1, P_2, and F_1 means are simply

$$\mu(P_1) = \mu_0 + \alpha_1^c - \delta_1^c \tag{11.4a}$$

$$\mu(P_2) = \mu_0 - \alpha_1^c - \delta_1^c \tag{11.4b}$$

$$\mu(F_1) = \mu_0 + \delta_1^c \tag{11.4c}$$

Rearranging and substituting observed for expected means, we obtain the three estimators,

$$\widehat{\mu}_0 = \frac{\bar{z}(P_1) + \bar{z}(P_2) + 2\bar{z}(F_1)}{4} \tag{11.5a}$$

$$\widehat{\alpha}_1^c = \frac{\bar{z}(P_1) - \bar{z}(P_2)}{2} \tag{11.5b}$$

$$\widehat{\delta}_1^c = \frac{2\bar{z}(F_1) - \bar{z}(P_1) - \bar{z}(P_2)}{4} \tag{11.5c}$$

With a model that includes all three forms of two-locus epistasis, there are six unknowns, so the mean phenotypes of at least six types of lines need to be evaluated. This is

Table 11.3 Coefficients for observed line means in expressions for estimated composite effects using the six-parameter model.

Parameter	P_1	P_2	F_1	F_2	B_1	B_2
μ_0	0	0	0	1	0	0
α_1^c	0	0	0	0	1	–1
δ_1^c	$\frac{1}{4}$	$\frac{1}{4}$	$-\frac{1}{2}$	2	–1	–1
α_2^c	0	0	0	–4	2	2
$\alpha_1^c\delta_1^c$	$\frac{1}{2}$	$-\frac{1}{2}$	0	0	–1	1
δ_2^c	$\frac{1}{4}$	$\frac{1}{4}$	$\frac{1}{2}$	1	–1	–1

most easily accomplished by assaying both parental lines, their F_1 and F_2 derivatives, and the two backcrosses of F_1 individuals to the parental lines (B_1 and B_2). As in the previous example, the estimated composite effects can then be written as simple linear functions of the observed means (Table 11.3). For example, the composite additive × additive effect is estimated by

$$\widehat{\alpha}_2^c = -4\bar{z}(\mathrm{F}_2) + 2\bar{z}(\mathrm{B}_1) + 2\bar{z}(\mathrm{B}_2)$$

Although we do not consider it in any detail here, it is worth noting that reciprocal crosses between parental lines can be used to estimate maternal effects and the effect of sex chromosomes. For example, when males of the P_1 line are crossed to females of the P_2 line, and vice versa, assuming that males are the heterogametic sex, the difference in mean phenotypes of daughters from the two lines provides an estimate of the difference between maternal effects associated with each parental line. In F_1 males, the effect of the X chromosome is superimposed on the maternal-effect difference. However, by making four possible types of crosses between F_1 individuals (two maternal sources of cytoplasm in females × two maternal sources of the X chromosome in males), a clean partitioning of these two sources of composite effects can be obtained (Carson and Lande 1984; Hard et al. 1992).

Because the estimates of the composite effects are linear functions of the line-cross means, the sampling variances of the estimates can be obtained from the expression for the variance of a sum (Equation 3.11b). Because the mean phenotype of each line is estimated independently of the others, there is no covariance between the different mean estimates. Thus, the sampling variance of a composite effect is simply the sum of the sampling variances of the line means used in its estimation, each weighted by the squared coefficient in the estimating equation. For example, for the preceding estimator of $\widehat{\alpha}_2^c$,

$$\mathrm{Var}(\widehat{\alpha}_2^c) = 16\mathrm{Var}[\bar{z}(\mathrm{F}_2)] + 4\mathrm{Var}[\bar{z}(\mathrm{B}_1)] + 4\mathrm{Var}[\bar{z}(\mathrm{B}_2)]$$

where $\mathrm{Var}[\bar{z}(\cdots)]$ is the squared standard error of a line mean.

Finally, we note in passing that one interesting application of line-cross methodology is in the analysis of selection response. In this setting, one grows saved seed from different generations (and often some of their crosses) from a selection experiment (or a breeding program) in a common environment and evaluates their means. This approach is discussed in WL Chapter 18 and is often called a **generation-means analysis** (**GMA**), and allows one to examine the component of selection response (Hammond and Gardner 1974; Smith 1979a, 1979b, 1983; Melchinger and Flachenecker 2006).

Hypothesis Testing

When the number of observations equals the number of parameters to be estimated, the solution of simultaneous equations cannot be used to evaluate the adequacy of the genetic model, as the parameter estimates are constrained to yield the observed line means exactly. This problem is eliminated when the number of observed line means exceeds the number of parameters to be estimated. For example, if data are available for the P_1, P_2, F_1, and F_2 lines, the statistic

$$\Delta = \bar{z}(\mathrm{F}_2) - \left(\frac{\bar{z}(\mathrm{P}_1) + \bar{z}(\mathrm{P}_2)}{4} + \frac{\bar{z}(\mathrm{F}_1)}{2} \right) \tag{11.6a}$$

provides a simple test for epistasis. In the absence of epistasis, the expected value of Δ is zero because at every locus, by the Hardy-Weinberg law, the F_2 is 25% $\mathrm{P}_1\mathrm{P}_1$, 50% $\mathrm{P}_1\mathrm{P}_2$, and 25% $\mathrm{P}_2\mathrm{P}_2$. The sampling variance of this test statistic is

$$\mathrm{Var}(\Delta) = \mathrm{Var}[\bar{z}(\mathrm{F}_2)] + \frac{\mathrm{Var}[\bar{z}(\mathrm{F}_1)]}{4} + \frac{\mathrm{Var}[\bar{z}(\mathrm{P}_1)] + \mathrm{Var}[\bar{z}(\mathrm{P}_2)]}{16} \tag{11.6b}$$

Under the reasonable assumption that the sampling distribution of Δ is approximately normal, the ratio $|\Delta|/\sqrt{\mathrm{Var}(\Delta)}$ provides a simple t test for epistasis. For large samples, if this ratio is greater than 1.96, then the null hypothesis of no epistasis can be rejected with 95% confidence.

A more general approach to hypothesis testing uses least-squares regression to estimate the model parameters and then compares the observed means with the model predictions. With this approach, one can start with a very simple model, evaluate its significance, and gradually add higher-order composite effects to the model, until no further significant improvement in the model fit occurs. This procedure, first suggested by Cavalli (1952) and Hayman (1960a), is known as the **joint-scaling test**.

Consider the simple additive model,

$$\bar{z}_i = \mu_0 + \theta_{Si}\alpha^c_i + e_i$$

where the *i*th line mean ($\bar{z}_i$) has coefficient θ_{Si}, and e_i denotes the deviation of the observed mean from the prediction of the model. In matrix form, letting $\bar{\mathbf{z}}$ be the vector of observed line means, $\mathbf{a}$ be the (2×1) vector of effects μ_0 and α^c_1, and $\mathbf{M}$ be the matrix of coefficients, the linear model becomes

$$\bar{\mathbf{z}} = \mathbf{Ma} + \mathbf{e} \tag{11.7}$$

where $\mathbf{e}$ is the column vector of residual errors, i.e., the vector of deviations between observed and predicted line means. Note that all of the elements in the first column of $\mathbf{M}$ are equal to one, as they are all multipliers for μ_0, whereas the second column contains the coefficients θ_{Si} for the various lines (Table 11.1).

Because the line means may vary with respect to accuracy (reflecting, for example, different sample sizes), they should not be weighted equally in the computation of $\mathbf{a}$. From Equation 10.13a, the weighted least-squares solution is

$$\widehat{\mathbf{a}} = (\mathbf{M}^T\mathbf{V}^{-1}\mathbf{M})^{-1}\mathbf{M}^T\mathbf{V}^{-1}\bar{\mathbf{z}} \tag{11.8}$$

where the covariance matrix $\mathbf{V}$ for the residuals is diagonal with diagonal elements equal to the squared standard errors of the means. The treatment of $\mathbf{V}$ as a diagonal matrix assumes that the measured individuals from the different lines are unrelated, i.e., that there is no sampling covariance between the observed means.

From Equation 10.13b, the sampling variances and covariances of the two parameter estimates, $\widehat{\mu}_0$ and $\widehat{\alpha}^c_1$, are given by the elements of the (2×2) covariance matrix

$$\mathrm{Var}(\widehat{\mathbf{a}}) = \mathbf{C} = (\mathbf{M}^T\mathbf{V}^{-1}\mathbf{M})^{-1} \tag{11.9}$$

The two diagonal elements of $\mathbf{C}$ are $\mathrm{Var}(\widehat{\mu}_0)$ and $\mathrm{Var}(\widehat{\alpha}^c)$, whereas both off-diagonal elements are equal to $\mathrm{Cov}(\widehat{\mu}_0, \widehat{\alpha}^c)$. Sampling covariance arises between estimates of the parameters because they are jointly estimated from a common data set. Letting $\widehat{z}_i = \widehat{\mu}_0 + \theta_{Si}\widehat{\alpha}^c$ be the fitted (predicted) mean phenotype for the *i*th line, the sampling variance of $\widehat{z}_i$ is simply the variance of a sum,

$$\mathrm{Var}(\widehat{z}_i) = \mathrm{Var}(\widehat{\mu}_0) + 2\theta_{Si}\mathrm{Cov}(\widehat{\mu}_0, \widehat{\alpha}^c) + \theta^2_{Si}\mathrm{Var}(\widehat{\alpha}^c) \tag{11.10a}$$

Using Equations 9.21b and 11.9, the covariance matrix for the predicted means is

$$\text{Var}(\,\widehat{\mathbf{z}}\,) = \text{Var}(\,\mathbf{M}\,\widehat{\mathbf{a}}\,) = \mathbf{MCM}^T = \mathbf{M}(\mathbf{M}^T\mathbf{V}^{-1}\mathbf{M})^{-1}\mathbf{M}^T \tag{11.10b}$$

Although the least-squares solution is completely general, significance testing requires the assumption of normality. If that assumption is met, the weighted error sum of squares

$$\chi^2 = \sum_{i=1}^{k} \frac{(\bar{z}_i - \widehat{z}_i)^2}{\text{Var}(\bar{z}_i)} \tag{11.11}$$

where k is the number of observed lines, provides a test statistic for the adequacy of the model (Appendix 3). Under the null hypothesis of purely additive gene action, this test statistic will be χ^2 distributed with degrees of freedom equal to the number of lines minus the number of estimated parameters, in this case giving $(k-2)$ df.

If the test statistic is large enough to reject the additive model, the logical next step is to evaluate the additive-dominance model. In this case, the vector $\mathbf{a}$ contains a third element, δ_1^c, and the matrix $\mathbf{M}$ contains a third row consisting of the elements θ_{Hi}. The solution again follows from Equation 11.8, and the new fit is evaluated by use of Equation 11.11, where the χ^2 statistic is now distributed with $k-3$ degrees of freedom, as three parameters are fitted.

Letting χ^2_A and χ^2_{AD} denote the test statistics associated with the additive and additive-dominance models, then the difference

$$\Lambda = \chi^2_A - \chi^2_{AD} \tag{11.12}$$

is equivalent to a likelihood-ratio test statistic (see Example A4.5). Such statistics are asymptotically χ^2 distributed (for large sample sizes), with degrees of freedom equal to the difference in the number of parameters included in the two models (in this case, one). While the inclusion of dominance in the model will definitely improve the fit, Equation 11.12 provides a test for whether the improvement is significant.

If the additive-dominance model is rejected on the basis of an overly large value for χ^2_{AD}, the next step is to proceed with the analysis of models containing epistatic effects, assuming that enough line means are available for such analysis. At this point, δ_1^c may or may not be dropped from the model depending on its degree of significance in the previous analysis. The significance of the improvement of fit with models containing epistasis can again be evaluated by use of the appropriate likelihood-ratio test, i.e., by the difference in χ^2 test statistics between the modified model and the previous restricted model.

Example 11.2. We now use the joint-scaling test to study the genetic basis of human skin color. The sample consists primarily of residents of Liverpool, England (Harrison and Owen 1964). Pigmentation was measured as the reflectance of the medial aspect of the right upper arm at 545 mμ. The P_1 consists of individuals of West African origin and the P_2 of individuals of European descent. The data are as follows

	P_1	P_2	F_1	F_2	B_1	B_2
$\bar{z}_i$	14.4	41.0	28.4	30.3	24.2	34.7
SE$(\bar{z}_i)$	0.611	0.453	0.581	1.483	1.334	1.122

There is a threefold range of variation in the standard errors of the mean phenotypes, and this translates into a tenfold range in the sampling variances. Clearly, weighted least-squares regression is desirable in this situation.

We start by considering the simplest genetic model, assuming that all gene action is additive within and between loci. Table 11.1 gives the coefficients for the effects μ_0 and α^c as

$$\mathbf{M} = \begin{pmatrix} 1 & 1 \\ 1 & -1 \\ 1 & 0 \\ 1 & 0 \\ 1 & 0.5 \\ 1 & -0.5 \end{pmatrix}$$

The sampling covariance matrix for the line means is diagonal with $V_{ii} = [\text{SE}(\bar{z}_i)]^2$, with

$$\mathbf{V} = \begin{pmatrix} 0.373 & 0.000 & 0.000 & 0.000 & 0.000 & 0.000 \\ 0.000 & 0.205 & 0.000 & 0.000 & 0.000 & 0.000 \\ 0.000 & 0.000 & 0.338 & 0.000 & 0.000 & 0.000 \\ 0.000 & 0.000 & 0.000 & 2.199 & 0.000 & 0.000 \\ 0.000 & 0.000 & 0.000 & 0.000 & 1.780 & 0.000 \\ 0.000 & 0.000 & 0.000 & 0.000 & 0.000 & 1.259 \end{pmatrix}$$

These lead to

$$\mathbf{M}^T\mathbf{V}^{-1}\mathbf{M} = \begin{pmatrix} 12.325 & -2.311 \\ -2.311 & 7.891 \end{pmatrix}$$

which yields the sampling covariance matrix for the parameter estimates (Equation 11.9),

$$\mathbf{C} = (\mathbf{M}^T\mathbf{V}^{-1}\mathbf{M})^{-1} = \begin{pmatrix} 0.086 & 0.025 \\ 0.025 & 0.134 \end{pmatrix}$$

and from Equation 11.8, the parameter estimates become

$$\widehat{\mathbf{a}} = \begin{pmatrix} \widehat{\mu}_0 \\ \widehat{\alpha}^c \end{pmatrix} = \begin{pmatrix} 28.17 \\ -13.07 \end{pmatrix}$$

The standard errors of the parameter estimates are equal to the square roots of the diagonal elements of $\mathbf{C}$,

$$\text{SE}(\widehat{\mu}_0) = (0.086)^{1/2} = 0.29$$
$$\text{SE}(\widehat{\alpha}^c) = (0.134)^{1/2} = 0.37$$

The line means predicted by the model are obtained as $\widehat{\mathbf{z}} = \mathbf{M}\widehat{\mathbf{a}}$, and the sampling variances and covariances of predicted values by Equation 11.10b,

$$\text{Var}(\widehat{\mathbf{z}}) = \mathbf{M}(\mathbf{M}^T\mathbf{V}^{-1}\mathbf{M})^{-1}\mathbf{M}^T$$

The square roots of the diagonal elements of this latter matrix are the estimated standard errors of the predicted means. In all cases, the predicted values are very close to the observed means:

	P_1	P_2	F_1	F_2	B_1	B_2
$\widehat{z}$	15.2	41.2	28.2	28.2	21.6	34.7
$\bar{z} - \widehat{z}$	−0.8	−0.2	0.2	2.1	2.6	0.0
SE($\widehat{z}$)	0.52	0.41	0.29	0.29	0.38	0.31

The test statistic, $\chi^2_A = 7.510$, with four degrees of freedom, is not significant as $\Pr(\chi^2_4 \geq 7.510) = 0.11$. Thus, the fitted model with $\widehat{\mu}_0 = 28.17$ and $\widehat{\alpha}^c = -13.07$ appears to adequately explain the data. Reevaluation of the data with the additive-dominance model confirms this conclusion. In this case, the analysis proceeds with

$$\mathbf{M} = \begin{pmatrix} 1 & 1 & -1 \\ 1 & -1 & -1 \\ 1 & 0 & 1 \\ 1 & 0 & 0 \\ 1 & 0.5 & 0 \\ 1 & -0.5 & 0 \end{pmatrix}$$

and

$$\mathbf{a} = \begin{pmatrix} \mu_0 \\ \alpha^c \\ \delta^c \end{pmatrix}$$

yielding the parameter estimates (and associated standard errors):

$$\widehat{\mu}_0 = 28.32\,(0.32), \quad \widehat{\alpha}^c = -13.14\,(0.37), \quad \widehat{\delta}^c = 0.44\,(0.34)$$

Notice that the estimates $\widehat{\mu}_0$ and $\widehat{\alpha}^c$ are very close to those obtained under the purely additive model, and that $\widehat{\delta}^c$ is only slightly greater than its standard error. The test statistic for this analysis is $\chi^2_{AD} = 5.879$. The likelihood-ratio test statistic, $\Lambda = \chi^2_A - \chi^2_{AD} = 1.631$, provides a test of the hypothesis that dominance accounts for a significant proportion of the variance among line means. With one degree of freedom, Λ is not significant, as $\Pr(\chi^2_1 \geq 1.631) = 0.20$.

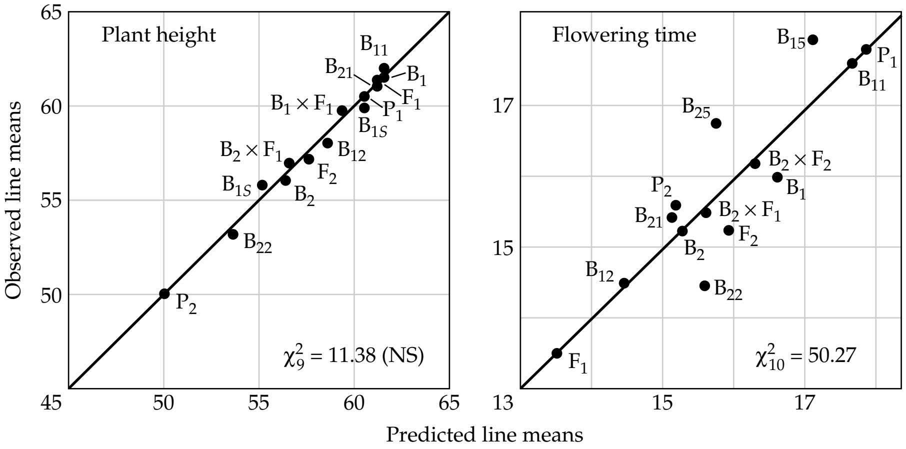


Figure 11.1 Observed versus fitted mean phenotypes for 14 line-cross derivatives from two pure parental lines of tobacco. The diagonal line gives the expected pattern if observed and predicted line means were identical. (After Jinks and Perkins 1969.)

Line Crosses in *Nicotiana rustica*

Few attempts have been made to estimate more than two-locus epistatic effects using line-cross analysis. However, Mather, Jinks, and their associates have done an enormous amount of work with highly inbred lines of tobacco to evaluate the relative importance of various types of higher-order gene action (Mather and Jinks 1982). An experiment by Jinks and Perkins (1969) is particularly noteworthy. In addition to the six fundamental line crosses, they created four second-generation backcrosses ($B_1 \times P_1$, $B_1 \times P_2$, $B_2 \times P_1$, $B_2 \times P_2$), the crosses $B_1 \times F_1$ and $B_2 \times F_1$, and the selfed backcrosses (B_{1s} and B_{2s}), and they assayed all

Table 11.4 Composite effects estimated from phenotypic means of two line crosses involving inbred parental stocks of tobacco, $V_1 \times V_5$ (upper lines of data for each trait) and $V_2 \times V_{12}$ (lower lines). Only significant effects are given, and the resulting models provide an adequate fit to the data in all cases. Leaf length was not analyzed in the $V_2 \times V_{12}$ experiment. (Data from Pooni et al. 1985.)

Character	$\widehat{\mu}_0$	$\widehat{\alpha}_1^c$	$\widehat{\delta}_1^c$	$\widehat{\alpha}_2^c$	$\widehat{\alpha}_1^c\widehat{\delta}_1^c$	$\widehat{\delta}_2^c$
Height—2 wk	5.50	1.19	1.10	1.27		
	4.34	1.97	0.48	0.55		
—4 wk	15.74	3.53	3.32	3.97		
	15.22	9.96	2.58	3.48		
—6 wk	49.84	5.02	−2.48	9.95		
	49.14	30.22	11.14	9.21	4.12	
—flowering	73.75	7.99		−5.02		
	83.40	5.84	3.12	−6.21		
—final	126.80	12.57	3.72	−7.58		
	143.44	8.79	21.91			−3.33
Leaf length	19.10	0.35		−1.89		

Table 11.5 A survey of the composite effects estimated in line-cross analyses. The results reported for each analysis describe the most parsimonious genetic model, and all recorded effects are statistically significant. The motivation for the use of logarithmic transformations in some cases is discussed in Chapter 14.

Character	$\widehat{\mu}_0$	$\widehat{\alpha}_1^c$	$\widehat{\delta}_1^c$	$\widehat{\alpha}_2^c$	$\widehat{\alpha}_1^c\widehat{\delta}_1^c$	$\widehat{\delta}_2^c$	Reference
Corn							
time to silking	65.19	0.20	−6.16	−4.40	7.33	5.92	Mohamed 1959
time to shed pollen	62.88	−1.57	−4.36	−1.94	4.74	3.69	
Lima beans							
seed size	0.57	−0.25	−0.16	0.04	0.30	0.19	Ryder 1958
Tomatos							
$\log_{10}$(fruit weight)	0.69	−0.87	0.02	0.12	−0.03		Powers 1951
Pitcher-plant mosquito							
$\log_{10}$(critical photoperiod)	−1.84	−4.45	0.40		1.79	4.18	Hard et al. 1992
Drosophila melanogaster							
ln(longevity)	1.79	−0.06	0.01	0.03			Luckinbill et al. 1988
Drosophila tripunctata							
ovipos. site pref.	0.31	−0.12	0.05	0.18			Jaenike 1987
D. heteroneura × *D. sylvestris*							
ln(head length)	3.00	0.09					Templeton 1977
ln(head width)	3.96	−0.01					
Astyanax (cave fish)							
eye diameter	4.72	−2.43	1.62	1.10	−1.57	−1.24	Wilkens 1971
Mice							
$\log_{10}$(body weight)	1.39	0.18	−0.01	−0.07	−0.01	0.01	Chai 1956
Chickens							
weight	3.49	−0.12	−0.06	−0.08	0.08	0.05	Waters 1931

of them simultaneously (i.e., in a common garden). With such a large number of lines, there are enough degrees of freedom to test for the significance of three-locus epistatic effects.

In the case of final plant height, a model with fitted parameters (and associated standard errors) $\widehat{\mu}_0 = 57.64 \pm 0.23$, $\widehat{\alpha}^c_1 = 5.32 \pm 0.20$, $\widehat{\delta}^c_1 = 3.55 \pm 0.38$, $\widehat{\alpha}^c_2 = 4.85 \pm 1.28$, and $\widehat{\alpha}^c_2\widehat{\delta}^c_1 = 3.73 \pm 1.03$ provided an excellent fit to the data (Figure 11.1). Thus, for this trait, there are significant additive × additive and additive × additive × dominance interactions involving genes from different lines, but no significant additive × dominance, dominance × dominance, or higher-order interactions. The most parsimonious three-locus fit for the data on flowering time is obtained with $\widehat{\mu}_0 = 15.91 \pm 0.09$, $\widehat{\alpha}^c_1 = 1.36 \pm 0.15$, $\widehat{\delta}^c_1 = -2.39 \pm 0.22$, and $\widehat{\alpha}^c_2\widehat{\delta}^c_1 = 1.81 \pm 0.37$. Here, there are significant differences between observed and predicted line means in this case (Figure 11.1), suggesting the existence of even higher-order epistatic interactions. Analyses with other parental lines (Smith 1937; Hill 1966) led to similar results—final height was usually described adequately by a model incorporating dominance and at least one form of two-locus epistasis, while flowering time was influenced by epistatic interactions between pairs, triplets, and higher numbers of loci. The results from two additional line-cross experiments, involving only the six fundamental generations, indicate that epistasis contributes to line differences in other characters in *Nicotiana* (Table 11.4). In these additional cases, however, effects involving more than pairs of loci need not be invoked to explain the data, because in no case are the observed means significantly different from the final model predictions.

Additional Data

A summary of results from some other line-cross studies is given in Table 11.5. In most cases, the parental lines are conspecific isolates known at the outset to differ (often substantially) in mean phenotypes. The main message is consistent with the *Nicotiana* results—differentiation of divergent lines almost always involves epistatic effects. We will see in Chapter 12 that epistatic interactions can sometimes be removed by a suitable scale transformation, but it is unlikely that this would be successful for all the tabulated studies.

VARIANCE OF LINE-CROSS DERIVATIVES

In principle, the joint-scaling test can be used to interpret the variances as well as the means obtained from line crosses. One of the most important applications of such an analysis is in the interpretation of the outbreaks of variation often seen in an F_2 generation (Figure 1.3). To keep things simple, let us assume that gene action is additive and that the environmental variance (here denoted as σ^2_E) is independent of the genetic background. The expected phenotypic variances for the parental lines and their F_1 offspring are then

$$\sigma^2(\mathrm{P}_1) = \sigma^2_E + \sigma^2_{A_1} \tag{11.13a}$$

$$\sigma^2(\mathrm{P}_2) = \sigma^2_E + \sigma^2_{A_2} \tag{11.13b}$$

$$\sigma^2(\mathrm{F}_1) = \sigma^2_E + \frac{1}{2}\sigma^2_{A_1} + \frac{1}{2}\sigma^2_{A_2} \tag{11.13c}$$

where $\sigma^2_{A_1}$ and $\sigma^2_{A_2}$ are the additive genetic variances in the P_1 and P_2 lines. The genotypic variance in the F_1 generation follows from the fact that, for each locus, F_1 individuals contain exactly one P_1 allele and one P_2 allele, and that the two haploid sets of alleles contribute variance $\sigma^2_{A_1}/2$ and $\sigma^2_{A_2}/2$.

An additional source of genetic variance will appear in any line-cross derivative for which there is variation among individuals in the proportion of P_1 and P_2 genes. For example, in the F_2 generation there is a 50% probability of being P_1P_2 and 25% probabilities of being P_1P_1 or P_2P_2 at any locus. This variation is in contrast to that in the F_1 generation where all individuals are P_1P_2. Letting $-\alpha^c_i/2$ and $+\alpha^c_i/2$ be the mean additive effects of alleles at the ith locus in the P_1 and P_2 lines, then the variance among F_2 individuals attributable to differences between parental lines at the locus is

$$\sigma^2_S(i) = 0.25(-\alpha^c_i)^2 + 0.5\left(\frac{\alpha^c_i}{2} - \frac{\alpha^c_i}{2}\right)^2 + 0.25(\alpha^c_i)^2 = \frac{(\alpha^c_i)^2}{2} \tag{11.14}$$

Summing over all loci, we obtain the **segregational variance,**

$$\sigma^2_S = \frac{1}{2}\sum_{i=1}^{n}(\alpha^c_i)^2 \tag{11.15}$$

which describes the excess variance that appears in the F_2 generation as a consequence of the segregation of parental-line genes. The expected variance in the F_2 generation is then

$$\sigma^2(F_2) = \sigma^2_E + \frac{1}{2}\sigma^2_{A_1} + \frac{1}{2}\sigma^2_{A_2} + \sigma^2_S \tag{11.16}$$

To obtain a general expression for the variances within backcross and more advanced generations, we take the following approach, again assuming additive gene action. Recall that if a line has associated parameters H and S, then the proportions of P_1-purebred, P_2-purebred, and crossbred genotypes at any locus are $S - (H/2)$, $1 - S - (H/2)$, and H, respectively. For the *i*th locus, these classes have genotypic means and variances equal to $(\alpha^c_i,\, \sigma^2_{A_1 i})$, $(-\alpha^c_i,\, \sigma^2_{A_2 i})$, and $(0,\, [\sigma^2_{A_1 i} + \sigma^2_{A_2 i}]/2)$, respectively. For any line, the genetic variance associated with the locus can be expressed as

$$\sigma^2_A(i) = E(A^2_i) - \mu^2_{A_i}$$

where A_i denotes the breeding value of an individual at the locus, and from arguments given above, $\mu_{A_i} = \theta_S\alpha^c_i = (1 - 2S)\alpha^c_i$. An expression for $E(A^2_i)$ is obtained by averaging over the three possible classes of genotypes,

$$\begin{aligned} E(A^2_i) &= \left(S - \frac{H}{2}\right)[\sigma^2_{A_1 i} + (\alpha^c_i)^2] + H\left(\frac{\sigma^2_{A_1 i} + \sigma^2_{A_2 i}}{2}\right) \\ &\qquad + \left(1 - S - \frac{H}{2}\right)[\sigma^2_{A_2 i} + (-\alpha^c_i)^2] \\ &= S\sigma^2_{A_1 i} + (1 - S)\sigma^2_{A_2 i} + (1 - H)(\alpha^c_i)^2 \end{aligned}$$

Putting these results together, for any derivative line with indices S and H, the variance associated with the *i*th locus is

$$\sigma^2_A(i) = S\sigma^2_{A_1 i} + (1 - S)\sigma^2_{A_2 i} + [4S(1 - S) - H](\alpha^c_i)^2 \tag{11.17}$$

Adding the environmental variance, summing over all loci, and recalling Equation 11.15, the expected phenotypic variance for a line with properties S and H is

$$\sigma^2 = \sigma^2_E + S\sigma^2_{A_1} + (1 - S)\sigma^2_{A_2} + 2[4S(1 - S) - H]\sigma^2_S \tag{11.18}$$

For example, for the B_1 backcross ($H = 1/2$ and $S = 3/4$) and the B_2 backcross ($H = 1/2$ and $S = 1/4$),

$$\sigma^2(B_1) = \sigma^2_E + \frac{3}{4}\sigma^2_{A_1} + \frac{1}{4}\sigma^2_{A_2} + \frac{1}{2}\sigma^2_S \tag{11.19a}$$

$$\sigma^2(B_2) = \sigma^2_E + \frac{1}{4}\sigma^2_{A_1} + \frac{3}{4}\sigma^2_{A_2} + \frac{1}{2}\sigma^2_S \tag{11.19b}$$

We are now equipped with all of the statistical machinery necessary to develop a predictive model for line variances. Analogous to the model developed above for line means, we let $\mathbf{v}$ be the vector of observed phenotypic variances for the various lines, $\mathbf{a}$ be the vector of variance components (σ^2_E, σ^2_{A1}, σ^2_{A2}, and σ^2_S), and $\mathbf{M}$ be the matrix of coefficients for the lines, with all elements in the first column being equal to one, and values of S, $(1-S)$, and $2[4S(1-S)-H]$ being entered in the remaining three columns. The linear model for line variances can then be summarized as

$$\mathbf{v} = \mathbf{Ma} + \mathbf{e} \tag{11.20}$$

where $\mathbf{e}$ is the vector of residual errors. The weighted least-squares estimates for the model parameters are given by

$$\widehat{\mathbf{a}} = (\mathbf{M}^T\mathbf{V}^{-1}\mathbf{M})^{-1}\mathbf{M}^T\mathbf{V}^{-1}\mathbf{v} \tag{11.21}$$

where $\mathbf{V}$ is the sampling covariance matrix for the line variances. Provided the individuals of the various lines are unrelated, then all of the elements of $\mathbf{V}$ are zero except those on the diagonal. These are set equal to $2v_j^2/(n_j+2)$, the unbiased estimator of the sampling variance of a variance under the assumption of normality (Equation A1.10c), where n_j is the sample size, and v_j is the observed phenotypic variance of the *j*th line.

There is one small remaining problem. Unless one is willing to assume that the parental lines are completely homozygous, i.e. $\sigma^2_{A1} = \sigma^2_{A2} = 0$, Equation 11.21 cannot be solved directly. The problem is that $\mathbf{M}$ is singular, because for any line, the coefficient for σ^2_E is equal to the sum of the coefficients for σ^2_{A1} and σ^2_{A2}. This difficulty can be circumvented by deleting the first column from $\mathbf{M}$ and reducing the variance component vector to $\mathbf{a}^T = [\sigma^2(\mathrm{P}_1)$, $\sigma^2(\mathrm{P}_2)$, $\sigma^2_S]$, where $\sigma^2(\mathrm{P}_1) = \sigma^2_E + \sigma^2_{A1}$ and $\sigma^2(\mathrm{P}_2) = \sigma^2_E + \sigma^2_{A2}$ are the phenotypic variances for the two parental lines.

Hayman (1960b) took this analysis a step further in producing a maximum-likelihood procedure. The diagonal elements of the matrix $\mathbf{V}$ have expectations equal to $2\sigma_j^4/(n_j+2)$, where σ^2_j represents the expectation of the appropriate entry in $\mathbf{v}$. Under the assumption that the additive model is correct, the projected least-squares values, $\widehat{\mathbf{v}} = \mathbf{M}\widehat{\mathbf{a}}$, should actually be better estimates of the within-line variances than the original elements of $\mathbf{v}$. This implies that the elements of $\widehat{\mathbf{a}}$ should be computed a second time by use of Equation 11.21 after substituting the elements of $\widehat{\mathbf{v}}$ into $\mathbf{V}$. This procedure is then iterated until the estimates of $\widehat{\mathbf{a}}$ stabilize. (Note that during the iterative process, it is only the elements of the covariance matrix $\mathbf{V}$, and not those of the vector of observed variances $\mathbf{v}$, that are modified recursively.)

Because $\mathbf{V}$ is diagonal with $V_{ii} = 2\,\sigma_j^4/(n_j+2)$, Equation A3.11a gives the χ^2 statistic for goodness of fit of the observed variances to the predictions of the additive model as

$$\chi^2 = \sum_{j=1}^{k} \frac{(v_j - \widehat{v}_j)^2}{2\,\widehat{v}_j^{\,2}/(n_j+2)} \tag{11.22}$$

where the $\widehat{v}_j$ are the final estimates of the σ^2_j, and the degrees of freedom associated with the test statistic equal the number of lines (k) minus three (the number of variance components estimated).

Example 11.3. We now apply the joint-scaling test for variances to Harrison and Owen's (1964) data on human skin color. Recall from Example 11.2 that the analysis of means supports the idea that this character has an additive genetic basis. On the other hand, the phenotypic variances, recorded in the following table, appear to be rather inconsistent with the additive model. For example, the F_1 variance is much higher than the average of the P_1 and P_2 variances, and even exceeds that of the F_2. However, because the sampling variance of a variance is quite large (as can be seen in the third row of the following table), there is some question as to the significance of these differences.

	P_1	P_2	F_1	F_2	B_1	B_2
v_j	14.918	21.098	31.748	26.382	37.366	37.766
n_j	40	103	94	12	21	30
$2v_j^2/(n_j+2)$	10.597	8.479	20.999	99.430	121.410	89.142

The coefficients for the variance components in the model, $\sigma^2(\mathrm{P}_1)$, $\sigma^2(\mathrm{P}_2)$ and σ_S^2, are obtained from Equation 11.18,

$$\mathbf{M} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0.5 & 0.5 & 0 \\ 0.5 & 0.5 & 1 \\ 0.75 & 0.25 & 0.5 \\ 0.25 & 0.75 & 0.5 \end{pmatrix}$$

Inserting the values from the table above, the initial sampling covariance matrix is

$$\mathbf{V} = \begin{pmatrix} 10.597 & 0.000 & 0.000 & 0.000 & 0.000 & 0.000 \\ 0.000 & 8.479 & 0.000 & 0.000 & 0.000 & 0.000 \\ 0.000 & 0.000 & 20.999 & 0.000 & 0.000 & 0.000 \\ 0.000 & 0.000 & 0.000 & 99.430 & 0.000 & 0.000 \\ 0.000 & 0.000 & 0.000 & 0.000 & 121.410 & 0.000 \\ 0.000 & 0.000 & 0.000 & 0.000 & 0.000 & 89.142 \end{pmatrix}$$

Substituting into Equation 11.21, we obtain the initial set of least-squares parameter estimates,

$$\widehat{\mathbf{a}} = (\mathbf{M}^T\mathbf{V}^{-1}\mathbf{M})^{-1}\mathbf{M}^T\mathbf{V}^{-1}\mathbf{v} = \begin{pmatrix} 18.120 \\ 23.681 \\ 14.441 \end{pmatrix}$$

The following table shows how the parameter estimates change over the next several rounds of iterations,

	Iteration						
Estimates	2	3	5	10	15	20	Final SE
Var(P_1)	22.199	22.930	23.446	23.583	23.586	23.586	4.207
Var(P_2)	25.994	25.505	25.185	25.103	25.102	25.102	3.166
Var(S)	17.163	17.038	16.876	16.830	16.829	16.829	10.411
χ^2	14.135	10.432	10.175	10.102	10.101	10.101	

The standard errors are the square roots of the diagonal elements of the final estimate of $(\mathbf{M}^T\mathbf{V}^{-1}\mathbf{M})^{-1}$. Using the final set of parameter estimates, the predicted line variances $(\widehat{v}_j)$, obtained from Equation 11.18, and their standard errors, obtained as the square roots of the diagonal elements of $\mathbf{M}(\mathbf{M}^T\mathbf{V}^{-1}\mathbf{M})^{-1}\mathbf{M}^T$, are

	P_1	P_2	F_1	F_2	B_1	B_2
$\widehat{v}_j$	23.586	25.102	24.344	41.172	32.379	33.137
SE	4.207	3.166	2.306	9.845	5.042	5.187

The final χ^2 value (10.101) is rather large, but because of possible nonnormality of the data, its statistical interpretation is somewhat questionable. Because the difference between observed and expected line variances is less than two standard errors $(2[\mathrm{SE}(v_j)^2 + \mathrm{SE}(\widehat{v}_j)^2]^{1/2})$ for all lines, there seems to be no strong justification for rejecting the additive model.

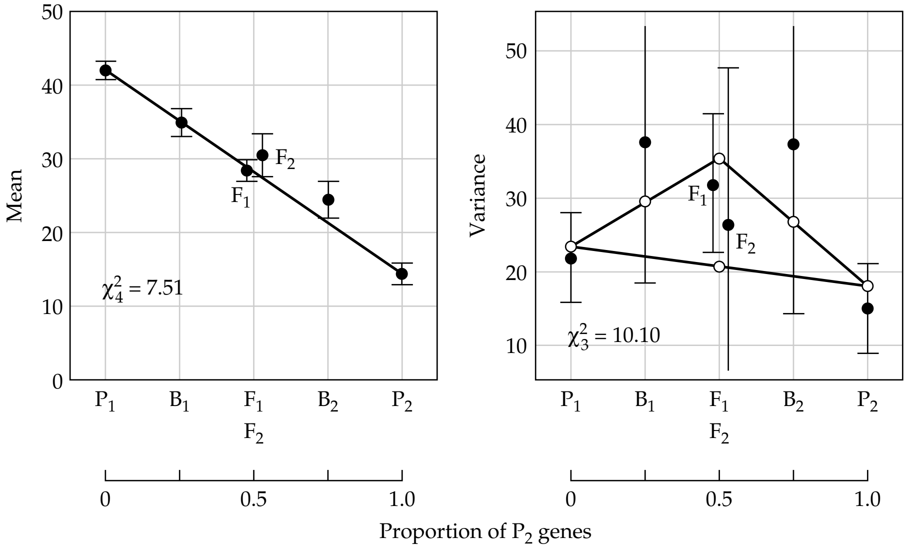


Figure 11.2 Observed means and variances (± twice the standard errors) of human skin color in relation to the predictions of the additive model. The maximum-likelihood predictions are given by the line on the left graph and the open circles on the triangle on the right graph. (From Harrison and Owens 1964.)

A graphical comparison of the observed and predicted means and variances of human skin color (Figure 11.2) serves to illustrate two important points. First, due to the large standard errors of variance estimates, *scaling tests based on variances are much less powerful than those based on means*. Thus, while the preceding methodology can be generalized to compute the dominance and epistatic components of the segregational variance (Mather and Jinks 1982), the statistical reliability of such analyses is very low, and we will not pursue it further.

Second, the additive model leads to some very simple geometric relationships (Figure 11.2). The expected line means are linear functions of the proportion of genes derived from each parental line. The expected line variances fall on a triangle, the vertices of which represent the expected P_1, P_2, and F_2 variances. The expected F_1 variance lies on the midpoint of the line leading from P_1 to P_2, while the expected B_1 and B_2 variances lie on the midpoints of the lines from P_1 to F_2 and P_2 to F_2, respectively.

BIOMETRICAL APPROACHES TO THE ESTIMATION OF GENE NUMBER

We now turn to a second application of line-cross analysis—estimation of the number of segregating loci responsible for quantitative variation. The subject is of importance for several reasons. First, since the beginning of the last century (Chapter 1), there has been considerable debate as to whether most large evolutionary changes are due to a small number of macromutations or to gradual substitution of minor allelic variants at a large number of loci (Gould 1980; Charlesworth et al. 1982; Gottlieb 1984; Coyne and Lande 1985; Orr and Coyne 1992; Wu and Palopoli 1994; WL Chapter 25). Second, from a more statistical point of view, the nice properties of normal theory that facilitate quantitative-genetic analysis become violated to a greater degree as the number of segregating loci becomes smaller. The ideal setting for much of quantitative-genetic theory is a very large number of loci, all with small effects, the so-called **infinitesimal model** (WL Chapter 24). Third, for situations in which most of the genetic variation is a function of one or two genes with major effects, a fine-scale Mendelian analysis (through the direct observation of segregation ratios) will often be possible, provided the environmental component of variance is not of overwhelming importance (Chapter 16).

There are two general approaches to estimating gene number. The **biometrical approach**, which is the subject of this chapter, is based on statistical properties (means and variances) of phenotype distributions, and uses these properties to indirectly infer the number of segregating factors that are likely to be responsible for them. The second approach involves the search for associations between segregating molecular markers and quantitative-genetic variation (Chapters 17–20). While these more direct molecular approaches have largely supplanted biometerical approaches, the latter are both of historical importance and still have utility. For example, Nandamuri et al. (2017) used this approach to estimate the number of genes involved in the regulation of opsin gene expression in Cichlid fish.

Although we are ultimately interested in the total number of loci (n) contributing to the variance in trait expression, most estimates of n are actually measures of the **effective number of factors** (n_e) by which the characters in two lines differ. This quantity is equivalent to the number of freely segregating loci with equal effects that would yield the observed pattern of line means and variances. It is important to realize that estimates of n_e do not include the potentially large fraction of loci that do not vary between lines, yet could lead to phenotypic differences given the right kinds of mutations. In addition, n_e cannot exceed the number of independently segregating chromosomal segments, i.e., the number of chromosomes plus the mean number of recombination events per gamete (the **segregation index**). In eukaryotes, there are usually one to two recombination events per chromosome. Thus, the maximum possible value of n_e is usually two to three times the haploid chromosome number, although each segregating unit can contain many loci.

In the following section, we will first describe an estimator, $\widehat{n}_e$, for the effective number of factors. We will then describe a more refined estimator, $\widehat{n}$, which provides estimates that are closer to the actual number of loci. Procedures for estimating n_e involve a number of assumptions, the most important of which is additivity of gene action. Thus, prior to analysis, a serious attempt should be made to find a scale on which the observed line means and variances are consistent with the additive model. The joint-scaling tests described on the previous pages provide an approach for testing the effectiveness of various scale transformations (see also Chapter 14).

The Castle-Wright Estimator

The most widely used method for estimating n_e utilizes information on the phenotypic means and variances of two parental lines and their line-cross derivatives (F_1, F_2, B_1, B_2, etc.) As first developed by Castle (1921) with his graduate student Sewall Wright (1968), the method was intended for use with inbred parental lines. Lande (1981) generalized it for use with genetically variable base populations, and the theory developed below is based on his modifications, and those of Cockerham (1986). In addition to additive gene action, the Castle-Wright technique assumes unlinked loci and equality of allelic effects, although we relax the latter assumption in the following derivation. It also assumes that all genes with a positive influence on the trait are sorted into one line and all those with negative influences into the other. Hence, such estimates are essentially of the *difference* in the net number of plus alleles that the lines differ by, and hence are underestimates of the true number of segregating differences.

Letting α_i^c be the composite additive effect for the *i*th locus, the mean phenotype of the P_1 line can be written as $\mu(\text{P}_1) = \mu_0 + \sum_{i=1}^n \alpha_i^c = \mu_0 + n\,\overline{\alpha_i^c}$, where $\overline{\alpha_i^c}$ is the average composite additive effect of a locus. That for the P_2 line is $\mu(\text{P}_2) = \mu_0 - n\,\overline{\alpha_i^c}$. Thus, the expected difference in mean phenotypes is $\mu(\text{P}_1) - \mu(\text{P}_2) = 2n\,\overline{\alpha_i^c}$. The segregational variance, defined in Equation 11.15, can also be written as

$$\sigma_S^2 = n[\sigma^2(\alpha_i^c) + (\overline{\alpha_i^c})^2]/2 \tag{11.23a}$$

where $\sigma^2(\alpha_i^c)$ is the variance of composite effects among loci. The trick to deriving an estimator of n_e is to note that the expected squared difference between parental line means is

$$[\mu(\mathrm{P}_1)-\mu(\mathrm{P}_2)]^2 = E\bigg[\bigg(2\sum_{i=1}^{n}\alpha_i^c\bigg)^2\bigg] = 4n\left[\,\overline{(\alpha_i^c)^2} + (n-1)(\,\overline{\alpha_i^c}\,)^2\,\right]$$

$$= 4n\left[\,n(\,\overline{\alpha_i^c}\,)^2 + \sigma^2(\alpha_i^c)\,\right] \qquad (11.23\text{b})$$

Taking the ratio of these two expressions yields

$$\frac{[\mu(\mathrm{P}_1)-\mu(\mathrm{P}_2)]^2}{\sigma_S^2} = \frac{8[\,n(\,\overline{\alpha}_i^c)^2 + \sigma^2(\alpha_i^c)\,]}{(\,\overline{\alpha}_i^c)^2 + \sigma^2(\alpha_i^c)}$$

which upon rearrangement yields

$$n_e = \frac{[\mu(\mathrm{P}_1)-\mu(\mathrm{P}_2)]^2(1+C_\alpha)}{8\sigma_S^2} - C_\alpha \qquad (11.23\text{c})$$

where $C_\alpha = \sigma^2(\alpha_i^c)/(\overline{\alpha_i^c})^2$ is the squared coefficient of variation of the locus-specific additive effects. In the very unlikely event that all of the assumptions of the model hold, then Equation 11.23c would define the actual number of loci. In recognition of the fact that one or more assumptions are likely to be violated, Equation 11.23c is denoted as a measure of the effective number of factors, n_e. In effect, Equation 11.23c states that the line means and segregational variance are distributed in the same way that would occur if the two populations were differentiated at n_e freely recombining loci with equal and additive effects.

In general, C_α is an unobservable quantity, but it must be positive. Ignoring this term for the time being, and substituting observed quantities for expectations, we obtain a biased estimator for the effective number of factors,

$$\widehat{n}_e = \frac{[\bar{z}(\mathrm{P}_1)-\bar{z}(\mathrm{P}_2)]^2 - \mathrm{Var}[\bar{z}(\mathrm{P}_1)] - \mathrm{Var}[\bar{z}(\mathrm{P}_2)]}{8\mathrm{Var}(S)} \qquad (11.24)$$

hereafter referred to as the **Castle-Wright estimator,** where the $\bar{z}(\mathrm{P}_i)$ and $\mathrm{Var}[\bar{z}(\mathrm{P}_i)]$ are, respectively, the observed means and sampling variances of the means for the *i*th parental line. Estimates of n_e in the literature often ignore the two Var terms in the numerator, but these are required to correct for the sampling error of the estimates of the line means (Cockerham 1986).

When data are available for the backcross generations, the segregational variance estimate, $\mathrm{Var}(S)$, can be obtained by the weighted least-squares procedure described above. In the absence of backcrosses, it can be computed as a linear function of the observed phenotypic variances within lines, either as

$$\widehat{\mathrm{Var}}(S) = \begin{cases} \mathrm{Var}(\mathrm{F}_2) - \mathrm{Var}(\mathrm{F}_1) & (11.24\text{b}) \\[2ex] \mathrm{Var}(\mathrm{F}_2) - \dfrac{2\mathrm{Var}(\mathrm{F}_1)+\mathrm{Var}(\mathrm{P}_1)+\mathrm{Var}(\mathrm{P}_2)}{4} & (11.24\text{c}) \end{cases}$$

The importance of using least-squares estimates whenever possible is seen from Example 11.3, where $\mathrm{Var}(\mathrm{F}_2) - \mathrm{Var}(\mathrm{F}_1)$ is negative, but the least-squares estimate of σ_S^2 is 17.

The large-sample variance of n_e, obtained from the equation for the variance of a ratio under the assumption of normality (Equation A1.19b), is approximately

$$\mathrm{Var}(\widehat{n}_e) \simeq \widehat{n}_e^2\left[\frac{4\{\mathrm{Var}[\bar{z}(\mathrm{P}_1)]+\mathrm{Var}[\bar{z}(\mathrm{P}_2)]\}}{[\bar{z}(\mathrm{P}_1)-\bar{z}(\mathrm{P}_2)]^2} + \frac{\mathrm{Var}[\mathrm{Var}(S)]}{[\mathrm{Var}(S)]^2}\right] \qquad (11.25\text{a})$$

If $\mathrm{Var}(S)$ is estimated by least-squares analysis, its sampling variance, $\mathrm{Var}[\mathrm{Var}(S)]$, is obtained directly from the matrix $(\mathbf{M}^T\mathbf{V}^{-1}\mathbf{M})^{-1}$, as described above. Otherwise, it is estimated by the sum of the variances of the variances used to compute $\mathrm{Var}(S)$, each weighted by the square of the appropriate coefficient. For example, if $\text{Var}(S)$ is estimated by $\text{Var}(\text{F}_2) - \text{Var}(\text{F}_1)$, then

$$\text{Var}[\text{Var}(S)] = \frac{2[\text{Var}(\text{F}_2)]^2}{n_{\text{F}_2} + 2} + \frac{2[\text{Var}(\text{F}_1)]^2}{n_{\text{F}_1} + 2} \tag{11.25b}$$

where n_{F_2} and n_{F_1} are the sample sizes. This follows because $\text{Var}(\text{F}_1)$ and $\text{Var}(\text{F}_2)$ are independent estimates, and because under the assumption of normality, the large-sample variance of a variance is $2\text{Var}^2/(n+2)$ (Equation A1.10c).

We noted above that failure to account for the variation in composite effects among segregating loci, i.e., ignoring C_α, will tend to depress $\widehat{n}_e$ below the true number of segregating loci. Additional factors will usually result in a further downward bias. For example, if the genes with positive effects are distributed among both parental lines, the difference between the parental line means will be less than the maximum value because the positive effects of genes at some loci will be canceled by negative effects at others. Consider the case of one line being fixed for $+1$ genes at locus 1 and -1 genes at locus 2 and another line being fixed for -1 genes at locus 1 and $+1$ genes at locus 2. The number of segregating loci in the F_2 generation is two, but because both parental lines have mean phenotypes equal to zero, the expected value of n_e yielded by Equation 11.24a equals zero. Such problems become immediately apparent when F_2 individuals exhibit phenotypes outside of the range of variation in both parental lines, a phenomenon known as **transgressive segregation** (Chapter 18), but the absence of such individuals does not rule out the possibility of transgression. In order to minimize such interpretative difficulties, most investigators utilize parental lines with the maximum range of variation between mean phenotypes. This is often accomplished by artificially selecting lines in the upward and downward direction for several generations prior to crossing.

By inflating the estimated segregational variance, linkage will also cause $\widehat{n}_e$ to be downwardly biased. Letting c_{ij} be the recombination fraction between loci i and j, a more general formula for the segregational variance in the F_2 generation is

$$\sigma_S^2 = \frac{1}{2}\left[\sum_{i=1}^{n}(\alpha_i^c)^2 + \sum_{i=1}^{n}\sum_{j\neq i}^{n}\alpha_i^c\alpha_j^c(1-2c_{ij})\right] \tag{11.26a}$$

where the term on the right is the disequilibrium covariance (generated by linkage creating an excess of parental gametes; Chapter 5). Assuming that the effects of pairs of alleles are uncorrelated with their map distances, then Equation 11.26a simplifies to

$$\sigma_S^2 = \frac{n}{2}\left[\sigma^2(\alpha_i^c) + (\,\overline{\alpha_i^c}\,)^2 + (n-1)(1-2\overline{c})(\,\overline{\alpha_i^c}\,)^2\right] \tag{11.26b}$$

In principle, the disequilibrium contribution to σ_S^2 can be removed by taking the F_2 generation through several additional generations of random mating, because this reduces the disequilibrium covariance by the factor $(1-c)$ each generation (Chapters 5 and 20). A modified estimate of the segregational variance can then be obtained as the difference between the variance in the advanced generation and that in the F_1 line.

An alternative way to deal with the problem of linkage is to use this more general expression for σ_S^2 given by Equation 11.26b, combined with our previous expression for the expected squared difference between line means to define the relationship between the effective number of factors and the actual number of loci. Substituting Equation 11.26b into 11.23c and rearranging leads to the expression

$$n = \frac{2\,\overline{c}\,n_e + C_\alpha(n_e - 1)}{1 - n_e(1-2\overline{c})} \tag{11.27}$$

which reduces to Equation 11.23c for the special case in which $\overline{c} = 0.5$. Zeng (1992) suggested that by substituting the estimate $\widehat{n}_e$, obtained by use of Equation 11.24a, into this expression, nearly unbiased estimates of the *actual* number of loci (n) are achievable.

In order to take advantage of Zeng's suggestion, we require, at the very least, estimates of $\bar{c}$ and C_α. We have already provided a number of estimates of $\bar{c}$ in Table 11.2, showing how these can be obtained from genetic maps of chromosomes under the assumption of randomly distributed loci. For many species, such detailed information is not available. However, provided the haploid chromosome number M is known, then a downwardly biased estimate of $\bar{c}$ is given by

$$\bar{c} = \frac{M-1}{2M} \tag{11.28}$$

which assumes that recombination only occurs between pairs of genes on different chromosomes (by independent assortment), and that all chromosomes contain equal numbers of genes. Fortunately, estimates of $\bar{c}$ using this approximation are not greatly different than the more refined estimates obtained by use of Equation 11.3. For example, for humans $(M = 23)$, maize $(M = 10)$, and *Arabidopsis* $(M = 5)$, Equation 11.28 yields estimates of $\bar{c} = 0.478$, 0.450, and 0.400, respectively, which contrast with the more exact computations, 0.490, 0.474, and 0.443. Thus, provided the minimum amount of information exists on the cytology of the organism, fairly reasonable estimates of $\bar{c}$ are achievable.

The squared coefficient of variation of effects, C_α, is much more elusive **< Newer references on distribution of effect sizes?>**. While estimates of the distribution of allelic effects are expected to be generated in the future as QTL mapping continues (Chapters 17–20), the only available estimates of this parameter derive from Keightley's (1994) analysis of data from mutation-accumulation experiments performed on lines of *Drosophila melanogaster* (Chapter 15). For abdominal bristle number, sternopleural bristle number, and viability, C_α is on the order of 6, 24, and 17, respectively. Unfortunately, aside from the fact that these estimates have very large sampling variances, it is unclear how similar the spectrum of effects of spontaneous mutations is to that of the effects of alleles normally segregating in natural populations. To the extent that they are representative, such high values of C_α suggest a very leptokurtic (L-shaped) distribution of allelic effects, with a very high density of small effects and a long tail to the right. For comparison, with a half-normal distribution (truncated at the mean), $C_\alpha = 0.57$, and with a negative exponential distribution, $C_\alpha = 1.0$.

The modifications suggested above are not trivial, as the magnitude of bias that variation in allelic effects and linkage causes with the Castle-Wright estimator can be quite large. Consider, for example, the situation in which $C_\alpha = 15$, the average of the results reported above, and suppose that the Castle-Wright estimator yields $\widehat{n}_e = 4$. Substituting into Equation 11.27, such an estimate in humans $(\bar{c} = 0.49)$ would be compatible with the presence of 53 actual loci, while for *C. elegans* $(\bar{c} = 0.42)$, it would imply the presence of 134 loci. Assuming constant allelic effects $(C_\alpha = 0)$, the estimate for humans would be essentially unbiased, as $\widehat{n}$ still equals four, but for *C. elegans*, $\widehat{n} = 9$.

By inflating the segregational variance in the F_2, nonadditive gene action is still another factor that has a downward influence on estimates derived by the Castle-Wright estimator. However, provided dominance is the primary source of nonadditivity, then the use of $2\text{Var}(\text{F}_2) - \text{Var}(\text{B}_1) - \text{Var}(\text{B}_2)$ as the estimate of the segregational variance in Equation 11.24a can eliminate most of the problem (Wright 1968; Ollivier and Janss 1993). Because the expectation of this quantity is identical to σ^2_S defined in Equation 11.26b, Equation 11.27 applies as well.

In summary, we find that violations of the various assumptions of the Castle-Wright model usually conspire to ensure that $\widehat{n}_e$ is an underestimate of the actual number of loci contributing to the divergence of lines. Although the bias can be very substantial (Zeng et al. 1990), most of it can be eliminated by making the modifications suggested above, i.e., by first computing $\widehat{n}_e$ by use of Equation 11.24a, then substituting this estimate and estimates of C_α and $\bar{c}$ into Equation 11.27, and solving. An approximate expression for the sampling variance of the improved estimate of the actual number of loci is given by

$$\text{Var}(\widehat{n}) = \frac{4\bar{c}^2(1+C_\alpha)^2\text{Var}(\widehat{n}_e)}{[1-\widehat{n}_e(1-2\bar{c})]^4} \tag{11.29}$$

(Zeng 1992), where Var$(\widehat{n}_e)$ is defined in Equation 11.25a. Simulations by Zeng (1992) suggest that $\widehat{n}$ provides much more reasonable estimates of the number of loci (n), than does $\widehat{n}_e$. However, the sampling variance of $\widehat{n}$ can be quite large. Even negative estimates are possible, when by chance the estimate of σ_S^2 is negative. (Negativity can occur with the Castle-Wright estimator as well.) Thus, any attempt to estimate n by either approach should be based on large sample sizes (ideally, with hundreds of individuals measured in each line).

Example 11.4. In previous examples involving human skin color, we found that $\bar{z}(\mathrm{P}_1) = 14.4$ and $\bar{z}(\mathrm{P}_2) = 41.0$. Squaring the standard errors of the means yields Var$[\bar{z}(\mathrm{P}_1)] = 0.205$ and Var$[\bar{z}(\mathrm{P}_2)] = 0.373$. The estimated segregational variance (Example 11.3) is Var$(S) = 17.264$, and its sampling variance is obtained by squaring its standard error, Var[Var$(S)] = 11.033^2 = 121.724$. Substituting into Equation 11.24a, we obtain $\widehat{n}_e = 5.1$. Substituting into Equation 11.25a, Var$(\widehat{n}_e) = 10.703$, giving the standard error of $\widehat{n}_e$ as $10.703^{1/2} \simeq 3.3$. Thus, the data suggest the hypothesis that the majority of the genetic difference in skin color between the major races of man is a consequence of a very small number of segregating factors. It should be kept in mind, however, that because of the low degree of accuracy of the estimated segregational variance, $\widehat{n}_e$ is a highly uncertain measure of the effective number of factors. Supposing, for heuristic purposes, that the estimate $\widehat{n}_e$ is accurate, what might the actual number of loci (n) contributing to the character be? From Table 11.2, we know that the mean recombination fraction for randomly distributed genes is extremely high in humans $(\bar{c} = 0.49)$. Substituting this and $n_e = 5$ into Equation 11.27, we obtain

$$n = \frac{4.9 + 4C_\alpha}{0.9}$$

Assuming that all loci have equal effects $(C_\alpha = 0)$, which seems unlikely, then $n = 5.4$. For $C_\alpha = 1,\ 10$, and 100, $n = 10,\ 50$, and 450. Thus, if the squared coefficient of variation of effects is much greater than one, the actual number of loci may greatly exceed the effective number of factors.

A survey of estimates of $\widehat{n}_e$ is given in Table 11.6. It should be emphasized that each estimate only applies to the specific pair of parental lines and that substantial differences would be likely if other parental stocks were used. Furthermore, the data are adequately described by an additive model in only a few cases, so most of the estimates are definitely biased in the downward direction by nonadditive gene action. Despite these limitations, while several of the analyses imply that a dozen or more loci are responsible for the differentiation of characters between parental lines, a number of cases suggest the possibility that a single major factor is involved. The latter conclusion may, of course, be substantially in error due to the approximate and biased nature of the biometrical approach. Nevertheless, the Castle-Wright model serves as a flag for situations in which a leading-factor (major gene) hypothesis (Chapter 15) warrants consideration.

Effect of the Leading Factor

If the assumptions of additive gene action and unlinked loci hold, then n_e provides some information on the effect of the **leading factor** (the locus accounting for the largest amount of the difference between parental means). Let $\phi_{max} = 2\alpha^c_{max}/[\,\mu(\mathrm{P}_1) - \mu(\mathrm{P}_2)]$ be the proportion of the difference between parental means attributable to the largest factor, and denote its effect by α^c_{max}. Equation 11.26a yields the inequality $\sigma_S^2 \geq (\alpha^c_{max})/2$, which upon substitution into Equation 11.23c gives

$$\phi_{max} \leq \sqrt{\frac{1 + C_\alpha}{n_e + C_\alpha}} \tag{11.30a}$$

as an estimate of the upper bound on the effect of the leading segregating factor. Because

Table 11.6 A sample of estimates (± their standard errors) of the effective number of segregating factors differentiating parental lines, obtained by use of Equation 11.24a. Whenever possible, the segregational variance was obtained by least-squares analysis. Under the additive model column + and − denote agreement and incompatibility with an additive model.

Species	Character	$\widehat{n}_e$	Additive Model	Reference
Corn	log(% oil + 1.87)	18 ± 2	+	Sprague and Brimhall 1949
	time to silking	1 ± 1	−	Mohamed 1959
	time to shed pollen	1 ± 1	−	
	ln (height)	4 ± 1	−	Emerson and East 1913
	ln (nodes)	5 ± 1	+	
	ln (internode length)	1 ± 1	−	
	ln (ear length)	13 ± 3	−	
	ln (seed weight)	13 ± 3	−	
Lima beans	seed size	17 ± 2	−	Ryder 1958
Red pepper	fruit shape	3 ± 1	−	Khambononda 1950
	fruit weight	13 ± 1	−	
Rice	plant height	1 ± 1	−	Mohamed and Hanna 1964
Goldenrod (*Solidago*)	date of anthesis	6 ± 2	−	Goodwin 1944
Nicotiana Langsdorffii × *N. Sanderae*	corolla length	13 ± 1	−	Smith 1937
Mimulus nasutus × *M. guttatus*	ln(flowering time)	1 ± 1	+	Fenster and Ritland 1994
	corolla width	2 ± 1	+	
	stamen level	3 ± 1	+	
Mimulus guttatus × *M. cupriphilus*	flower width	5 ± 2	+	Macnair and Cumbes 1989
	flower height	4 ± 1	+	
	pistil length	18 ± 18	+	
	corolla length	6 ± 3	+	
Tomato	$\log_{10}$ (fruit weight)	12 ± 1	−	Powers 1942
Pearl millet (*Pennisetum*)	height	4 ± 1	−	Burton 1951
		4 ± 1	−	
	internode length	2 ± 1	−	
	leaves/stem	5 ± 1	−	
		7 ± 1		
Drosophila melanogaster	ln (longevity)	1 ± 1	−	Luckinbill et al. 1988
Drosophila tripunctata	ovipos. site pref.	1 ± 1	−	Jaenike 1987
Drosophila heteroneura × *D. silvestris*	head length	7 ± 4	+	Templeton 1977
	head width	1 ± 1	+	
Cave fish (*Astyanax*)	eye diameter	6 ± 1	−	Wilkens 1971
Chickens	weight	5 ± 1	−	Waters 1931
Mice	$\log_{10}$ (weight)	12 ± 1	−	Chai 1956

each segregating factor contains one or more loci, ϕ_{max} is also an upper bound on the effect of the leading locus. Lander and Botstein (1989) proposed a simple idea that yields a lower bound estimate for the effect of the leading factor. Under the assumption that one strain contains all "positive" genes and the other all "negative" genes (which might be approximated if the two strains were obtained by intense selection in opposite directions from a common stock), there must be at least one segregating factor with an effect at least as great as $[\,\mu(\text{P}_2) - \mu(\text{P}_1)]/n_e$. Expressed in terms of the proportion of the total difference, this effect is simply

$$\phi_{max} \geq 1/n_e \tag{11.30b}$$

Alternatively, from the standpoint of individual loci, an estimate of the minimum effect of the leading locus is $1/n$. Finally, because (from Equation 11.27) there are at least $2\bar{c}n_e$ loci, the maximum value of the *average* allelic effect is

$$\bar{\phi}_{max} = 1/(2\bar{c}n_e) \tag{11.30c}$$

Estimates of the statistical bounds on the effects of leading factors are of practical importance, because they can provide insight into the likely utility of molecular marker-based searches for loci with major effects (Chapters 17–20). Unfortunately, none of the above-mentioned statistics is particularly informative in this regard. Even when they are reliable, large estimates of the upper bound of the leading factor do not necessarily imply that any locus actually has a large effect, and although a large ϕ_{min} implies that at least one locus has a major effect, a small ϕ_{min} does not rule out the possibility of several loci with major effects. Further, as we detail in Chapter 20, a small GWAS effect *does not* imply a small allelic effect, as GWAS effect sizes are usually based on variances. An allele of large effect that is rare can have a small variance, and (as detailed in Chapter 20), an inverse relationship between allele effect size and allele frequency is often seen, with large-effect alleles tending to be rare (at low frequencies).

Zeng (1992) suggested an alternative approach to predicting the effects of leading factors. Given an estimate of C_α, one first derives the estimate of the number of loci, $\widehat{n}$. This provides an estimate of the mean allelic effect as

$$\widehat{\alpha} = [\bar{z}(\mathrm{P}_1) - \bar{z}(\mathrm{P}_2)]/\widehat{n} \tag{11.31}$$

Together, $\widehat{\alpha}$ and C_α then provide the first two moments of the distribution of allelic effects. If the form of the distribution is specified and uniquely defined by the mean and variance, one can then draw $\widehat{n}$ random effects from the distribution, order them, and evaluate the relative contributions of the various factors to the parental line divergence. As noted above, even slightly different values of C_α can lead to very different estimates of n. However, Zeng (1992) found that the number of significant loci (for example, the number that account for 90% of the total divergence) is extremely insensitive to changes in C_α, changing by only five or so over a range in which n changes by hundreds. Thus, for large surveys in which highly reliable estimates of phenotypic means and variances can be acquired, Zeng's approach has promise as a means of estimating the number of major factors.

Example 11.5. Here we present an alternative analytical approach for estimating the number of leading factors and their effects. This approach assumes that something is known about the form of the distribution of allelic effects. Let $p(\alpha)$ be the probability density function of the effects, α_i, and let $F(\alpha)$ be the cumulative frequency distribution, the probability that the effect of a randomly drawn gene is less than α. By definition, $dF(\alpha)/d\alpha = p(\alpha)$. Suppose now that n genes are randomly drawn from the distribution $p(\alpha)$, and rank ordered in terms of increasing effect, such that α_1 is the smallest effect and α_n is the largest effect (the leading factor). From the perspective of genetic analysis, one would like to know the expected effects of α_n, α_{n-1}, α_{n-2}, and so on. The theory of **order statistics** (Harter 1961; Sarhan and Greenberg 1962; Harter 1970a, 1970b; Kendall and Stuart 1977; David 1981) provides a potential solution.

Consider the *r*th smallest value in the set of random draws of n genes. The probability that at least r draws in a sample do not exceed the value α is

$$F_r(\alpha) = \sum_{i=r}^{n} \binom{n}{i} [F(\alpha)]^i [1 - F(\alpha)]^{n-i} \tag{11.32a}$$

which leads to the probability density function for the *r*th order statistic,

$$p_r(\alpha) = \frac{dF_r(\alpha)}{d\alpha} = \frac{n!}{(r-1)!(n-r)!} [F(\alpha)]^{r-1} [1 - F(\alpha)]^{n-r} p(\alpha) \tag{11.32b}$$

Thus, the expected value of the *r*th smallest factor is given by

$$E[\alpha_r] = \int_{-\infty}^{+\infty} \alpha p_r(\alpha) d\alpha \tag{11.32c}$$

which, for the leading factor, reduces to

$$E[\alpha_n] = n \int_{-\infty}^{+\infty} \alpha [F(\alpha)]^{n-1} p(\alpha) d\alpha \tag{11.32d}$$

An estimate of the proportional contribution of the leading factor to the line differentiation is $2E(\alpha_n)/[\bar{z}(\mathrm{P}_1) - \bar{z}(\mathrm{P}_2)]$.

These expressions only outline a general approach. Their actual implementation requires that one define the form of the probability density function $p(\alpha)$ (for example, a normal or a gamma distribution), and then characterize the function in terms of its parameters (usually the mean and variance of effects). Once an estimate of the number of loci (n) has been acquired (for example, by the use of Equation 11.27), these parameters are specified. The mean effect is estimated by $\bar{\alpha} = [\bar{z}(\mathrm{P}_1) - \bar{z}(\mathrm{P}_2)]/2n$, and the variance is defined by $\bar{\alpha}^2 C_\alpha$.

Table 11.7 Expected means and variances for line crosses derived from two haploid parental lines, under the assumption of zero linkage and additive gene action. Composite additive effects are defined as haploid effects, so that parental line divergence is still $2\alpha_1^c$ as in the diploid model.

Line	Mean	Variance
P_1	$\mu_0 - \alpha_1^c$	σ_E^2
P_2	$\mu_0 + \alpha_1^c$	σ_E^2
F_1	μ_0	$\sigma_E^2 + \sigma_S^2$
F_2	μ_0	$\sigma_E^2 + \sigma_S^2$
B_1	$\mu_0 - 0.5\alpha_1^c$	$\sigma_E^2 + \frac{3}{4}\sigma_S^2$
B_2	$\mu_0 + 0.5\alpha_1^c$	$\sigma_E^2 + \frac{3}{4}\sigma_S^2$

Extension to Haploids

The Castle-Wright model can be extended to the estimation of gene number in haploid organisms without great difficulty (Chovnick and Fox 1953). Because most haploids can be maintained clonally, we assume that the environment is the sole source of variation within the parental lines. The expected means and variances of the derived generations are laid out in Table 11.7. Note that the F_1 generation exhibits the same segregational variance as the F_2 due to the complete segregation of parental genes following fertilization and the production of haploid progeny. This segregational variance is

$$\sigma_S^2 = \sum_{i=1}^{n} 0.5[(\alpha_i^c - 0)^2 + (-\alpha_i^c - 0)^2] = n[\sigma^2(\alpha_i^c) + (\overline{\alpha_i^c})^2] \tag{11.33}$$

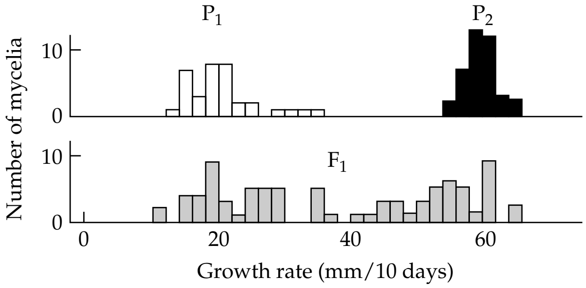
which is twice the expectation in the case of diploidy. Thus, the estimation equation for n_e with haploid organisms is the same as in the case of diploidy except that $4\mathrm{Var}(S)$, rather than $8\mathrm{Var}(S)$, appears in the denominator of Equations 11.23c and 11.24a. The equation for the sampling variance for $\widehat{n}_e$ needs to be multiplied by four, but the estimators for gene number, Equations 11.27 and 11.29, still apply provided the segregational variance is estimated in the F_1 generation.

Table 11.8 The proportion of deviant derived lines that are expected to contain a single gene from the donor parent (last column). (Data from Wehrhahn and Allard 1965.)

n	k	p_k	$p(1)$	$p(r \geq 1)$	$p(1)/p(r \geq 1)$
4	2	$\frac{1}{8}$	0.335	0.414	0.809
	3	$\frac{1}{16}$	0.206	0.227	0.905
	4	$\frac{1}{32}$	0.114	0.119	0.953
10	2	$\frac{1}{8}$	0.376	0.624	0.510
	3	$\frac{1}{16}$	0.350	0.650	0.735
	4	$\frac{1}{32}$	0.235	0.765	0.863

A very similar strategy can be employed with **doubled-haploid lines**. Such lines, which can be produced by a variety of cytological techniques (Kermicle 1969; Nitzsche and Wenzel 1977; Choo 1981), are homozygous at all loci. If two such parental lines are crossed to produce an F_1 generation, and a random sample of F_1 gametes is used to produce a new series of doubled haploids, the effective number of factors differentiating any two lines can be obtained by computing the segregational variance as half the difference between the F_1 doubled-haploid variance and the average variance in the P_1 and P_2, and employing Equation 11.24a (Choo and Reinbergs 1982). (The segregational variance is inflated twofold by the enforcement of homozygosity at each locus in doubled haploids.) The most reasonable estimate of n_e obtained with this approach utilizes parental lines with the highest and lowest mean phenotypes in a random sample from the population. However, with even a moderate number of segregating loci, the probability of obtaining the two most extreme lines possible is low unless the number of lines assayed is very large. Choo and Reinbergs (1982) used this technique to show that at least eight segregating factors contribute to the variation i grain yield, heading date, and plant height in barley. For additional biometrical approaches to gene number estimation in doubled haploids, see Snape et al. (1984). Further extension of the Castle-Wright approach were offered by Comstock and Enfled (1981) who allow for genes to act in a multiplicative (as opposed to additive) fashion, and by Jones (2001) who considered sex linkage and haplodiploidy.

Example 11.6. Croft and Simchen (1965) isolated dikaryotic mycelia from wild populations of the fungus *Collybia velutipes* and from these extracted asexually and sexually derived monokaryotic (haploid) spores. (A dikaryotic mycelium is a filament comprised of fused cells of two different parental origins, each containing a haploid nucleus). The growth rates of germinating spores were then assayed on a laboratory medium. Barring mutations, the growth rate of each asexual propagule is expected to be representative of one of the parental lines, because these propagules contain a single, nonrecombinant nucleus. On the other hand, the sexually derived progeny will exhibit segregational variance. Frequency distributions are given below for both types of offspring for one particular isolate.



The mean growth rates of the two parental types differ by $\bar{z}(\mathrm{P}_1) - \bar{z}(\mathrm{P}_2) = 39.45$ (mm/10 days), and the sampling variances of the two means are Var$[\bar{z}(\mathrm{P}_1)] = 2.43$ and Var$[\bar{z}(\mathrm{P}_2)] = 0.53$. The variance of growth rate among haploid replicates is 26.80, while the excess variance among sexual propagules is 277.80. Taking the latter quantity to be an estimate of the segregational variance, Var(S), substitution into Equation 11.24a (multiplied by 2) yields $\widehat{n}_e = 1.4$. The standard error is approximately 0.5. These results are reasonably consistent with those obtained from four other isolates: $1 \pm 0.2, 5 \pm 0.8, 3 \pm 0.3,$ and 1 ± 0.1. Thus, it seems likely that most of the growth rate differences among parental strains may be attributable to one to three loci.

OTHER BIOMETRICAL APPROACHES TO GENE NUMBER ESTIMATION

In outlining the Castle-Wright model, we emphasized several assumptions, violations of which tend to result in underestimation of the actual number of segregating loci. Although two potential problems, gametic phase disequilibrium and dominance, appear to be reconcilable, two others are less tractable—transgressive segregation (both lines contain plus and minus alleles) and variation among loci for allelic effects. We now consider three approaches that have been developed to circumvent these problems, all involving the use of species that can be self-fertilized.

The Inbred-backcross Technique

Wehrhahn and Allard (1965) developed a useful technique that yields estimates of both the minimum number of genetic factors responsible for the differentiation of two lines and the magnitude of the locus-specific effects. Two pure lines are crossed to form F_1 individuals, each of which is then backcrossed to one of the parental populations (say the P_1) for k generations (i.e., an $\mathrm{F}_1 \times \mathrm{P}_1$ cross, followed by a cross of their progeny to the P_1, etc.) The backcross descendants are then inbred for several generations to fix any segregating factors. The rationale for this breeding scheme is that as k becomes large, the probability that any inbred backcross line will retain more than one allele from the donor parent (the P_2) becomes small. The effects of individual genes can then be ascertained by comparing the phenotypic means of the recurrent parent (P_1) and the derived lines. A unique advantage of the inbred-backcross technique is that, after single-gene deviant lines have been isolated, lines with pairs of various genes can be constructed to evaluate epistatic effects between specific isolated factors. Likewise, genotype × environment interaction involving individual genes can be examined by growing lines in different environments.

A more formal statement of these arguments is as follows. The probability that a specific gene from the P_2 is incorporated into a specific line after k generations of backcrossing is $p_k = (1/2)^{k+1}$. If n freely segregating factors are responsible for the character difference between the P_1 and P_2, then the probability that a derived line retains just one of the P_2 alleles is

$$p(1) = np_k(1 - p_k)^{n-1} \tag{11.34a}$$

This follows directly from the binomial distribution, because each gene is retained or lost independently with probability p_k. The probability that a derived line contains at least one P_2 gene is

$$p(r \geq 1) = 1 - (1 - p_k)^n \tag{11.34b}$$

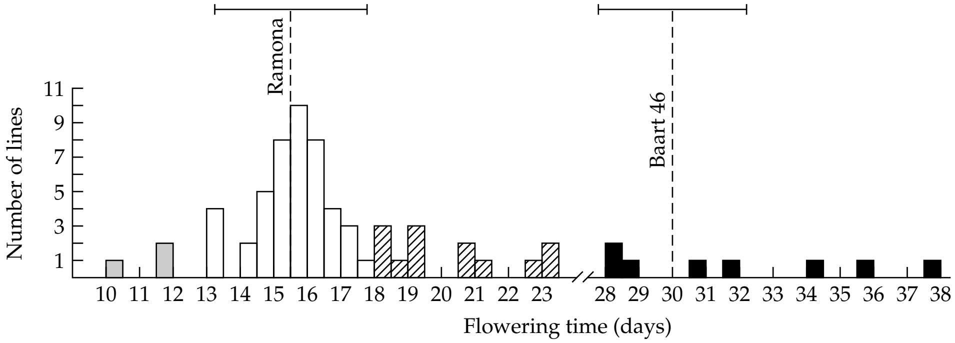
The ratio $p(1)/p(r \geq 1)$ is the conditional probability that any deviant derivative line contains only a single P_2 gene. Table 11.8 shows that unless n is very large, this probability is very high after only three or four generations of backcrossing.

If, by statistical comparison of phenotypic means, a fraction $\widehat{p}(r \geq 1)$ of the inbred-backcross lines is observed to differ significantly from the P_1, then a minimum estimate of the number of effective factors can be found by rearrangement of Equation 11.34b,

$$\widehat{n}_{WA} = \frac{\ln[1 - \widehat{p}(r \geq 1)]}{\ln(1 - 0.5^{k+1})} \tag{11.35}$$

(Mulitze and Baker 1985a, 1985b). Unlike the Castle-Wright estimator, Equation 11.35 yields estimates of n_e that are essentially independent of the degree of transgression of gene effects in the P_1 and P_2, i.e., a gene with a low effect in an otherwise high-performing line can be detected when substituted into a low-performing line.

Example 11.7. Wehrhahn and Allard (1965) crossed two pure lines of wheat (Ramona and Baart 46) with very different heading dates (flowering times). Two successive backcrosses were made to Ramona, followed by three generations of selfing, to produce 69 inbred backcross lines. How many lines are expected to contain any specific Baart 46 gene? Because $k = 2$, we have $p_k = (1/2)^3 = 1/8$, and because there are 69 total lines, $69/8 = 8.6$ of these are expected to carry a Baart 46 gene at a specified locus. From the properties of the binomial distribution, the standard error of the estimate is $[69p_k(1 - p_k)]^{1/2} = 2.8$. The distribution of heading date in the derived lines shows three groups of deviants from the Ramona distribution: (1) a group of eight very late lines that appear to contain a factor with major effect (black bars), (2) a group of 14 lines with slightly late heading dates (striped bars), and (3) a group of three lines with earlier heading dates (gray bars). The means and 95% confidence limits for the parental lines are given by the vertical and horizontal lines, and the three groups of deviants from the Ramona (recurrent) line are differentially shaded.



By other means, the authors showed rather convincingly that the group of 14 was heterogeneous for two factors. Thus, the difference in heading date between the two parental lines is caused by at least four effective factors, one of which operates in a direction opposite to the others. These four factors accounted for 96% of the line differentiation (80% was due to the leading factor), so if additional loci are involved, their effects must be very small. To see that the observations in the figure are consistent with Equation 11.35, let $\widehat{p}(r \geq 1) = (8 + 14 + 3)/69 = 0.362$. We then obtain $\widehat{n}_{WA} = 3.4$, which rounds up to 4.

Genotype Assay

Jinks and Towey (1976; Towey and Jinks 1977) developed a method for estimating n_e that is similar in philosophy to the inbred-backcross technique. In this case, however, the F_1 progeny of a cross between two pure lines are self-fertilized (instead of backcrossed). Their descendants are also self-fertilized until generation F_k, where k is usually 2 to 5. Two random (selfed) progeny are then raised from each F_k individual and selfed, and their offspring are assayed in a randomized design (Figure 11.3). A comparison of means (t test) and variances (F test) between the two families is used to detect whether one or more loci were segregating in the F_k grandparent. Not all heterozygotes are detected by this

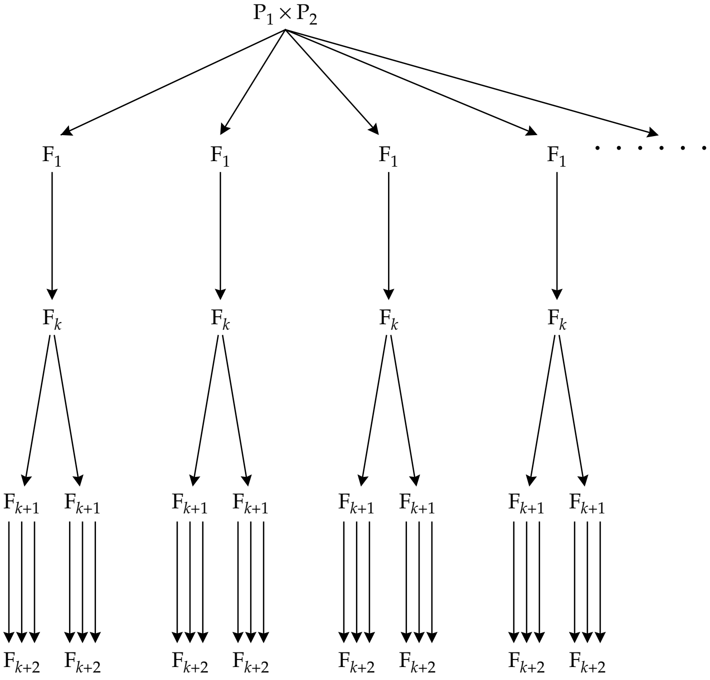


Figure 11.3 The crossing scheme involved in the genotype assay technique of Jinks and Towey (1976). Progeny from all generations beyond the F_1 are obtained by selfing. In generation $k+2$, multiple progeny are assayed from each of two sublines for each of the isolated selfed lineages.

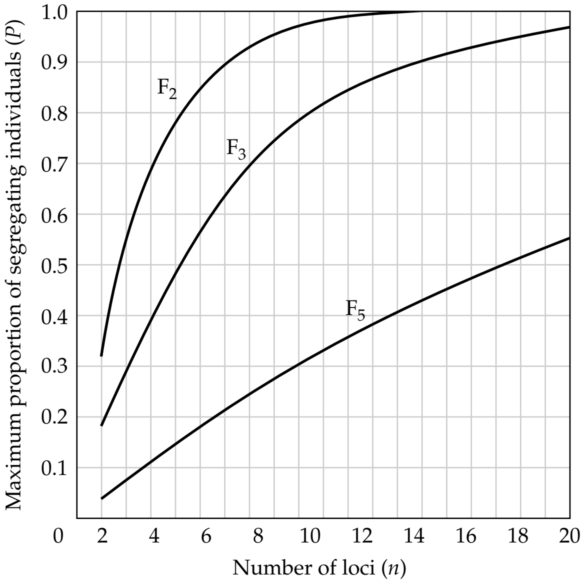


Figure 11.4 The expected fraction of individuals in the F_2, F_3, and F_5 generations detected as segregating by a genotype assay on their grandchildren, under the assumptions leading to Equation 11.36. (From Jinks and Towey 1976.)

approach, because the probability that two randomly chosen offspring of a heterozygous parent will differ in genotype is only 5/8 (the probability that one is BB and the other is Bb is $2 \times (1/4) \times (1/2) = 1/4$, that one is BB and the other is bb is $2 \times (1/4) \times (1/4) = 1/8$, and that one is Bb and the other is *bb* is $2 \times (1/2) \times (1/4) = 1/4$). However, the observed fraction of segregating grandparents can still be used to make inferences about n.

The logic behind this idea is as follows. Under continuous self-fertilization, heterozygosity is reduced by 50% each generation, so that the probability that an F_k individual is a heterozygote at a particular locus is $(1/2)^{k-1}$. The probability that two assayed progeny from a random F_k individual differ at this locus is then $(5/8) \cdot (1/2)^{k-1} = 5/2^{k+2}$. It follows that the probability that the two descendant sublines differ at least at one segregating locus is

$$P = 1 - \left(1 - \frac{5}{2^{k+2}}\right)^n \tag{11.36}$$

Rearranging and substituting observations for expectations, we obtain another estimator for the effective number of segregating factors,

$$\widehat{n}_{JT} = \frac{\ln(1 - \widehat{P})}{\ln\left(1 - \dfrac{5}{2^{k+2}}\right)} \tag{11.37}$$

There are several potential causes of bias in this estimator, most of which will lead to the usual underestimation of the actual number of loci. First, as in the inbred-backcross technique, the observed fraction of intrapair differences $(\widehat{P})$ is a matter of statistical power, increasing with the size of the assayed families, but decreasing with more stringent criteria for statistical significance. Second, the number of segregating factors will be depressed below the actual number of loci by linkage, the magnitude of this bias decreasing with increasing k (more opportunity for recombination). Hill and Avery (1978) consider this issue in some detail. Third, dominance, epistasis, and gametic phase disequilibrium can cause the expected phenotypes associated with different genotypes to be the same. Taking this masking problem into consideration, Jinks and Towey (1976) and Mulitze and Baker (1985a) have derived an alternative to Equation 11.37 that yields an upper (rather than lower) bound to n_e.

The relationship between P and n_{JT} varies rather substantially with the number of generations (k) prior to the genotype assay (Figure 11.4). If there are only a few effective factors, the greatest sensitivity is achieved (i.e., there is a strong response of P to n) when $k = 2$, provided linkage is unimportant. However, if there are more than five segregating loci, an F_2 assay is of little use, whereas an F_5 assay is quite sensitive.

Towey and Jinks (1977) applied the genotype assay to five generations of a cross between two lines of *Nicotiana rustica*. Even after six generations of selfing, there was no obvious decline in the fraction of descendant pairs exhibiting variation, for either flowering time or plant height (Table 11.9). Consequently, the estimates of n_{JT} for both characters increased approximately tenfold throughout the study.

The authors interpreted this increase to be a consequence of the gradual elimination of gametic phase disequilibrium through progressive rounds of recombination. If nothing else, this interpretation justifies our earlier discussion about the bias caused by linkage.

Panse's Technique

This infrequently utilized technique also starts with a cross between two pure lines. The F_1 individuals are self-fertilized to produce an F_2 generation, the members of which are selfed to produce a series of F_3 families. For any segregating locus, half of the F_2 individuals are expected to be heterozygous and the remaining half homozygous. Assuming additive effects, it follows from Chapter 4 that the genetic variance resulting from the *i*th locus in a population in Hardy-Weinberg equilibrium is $2pq\alpha_i^2$. This genetic variance is $\alpha_i^2/2$ for an F_3 family descending from a heterozygote, and zero otherwise. Thus, the average genetic variance within F_3 familes is $\sum_{i=1}^n \alpha_i^2/4$, or $\overline{\sigma_A^2} = n\alpha^2/4$ for n loci with equal effects.

Table 11.9 Minimum number of effective factors responsible for the differentiation of two lines of tobacco (*Nicotiana rustica*) as determined by genotype assay in progressive generations (F_k). n_{JT} is computed with Equation 11.37 and rounded up to the nearest unit. (Date from Towey and Jinks 1977.)

Grandparent Generation:	F_2	F_3	F_4	F_5	F_6
Flowering time					
$\widehat{P}$	0.200	0.450	0.314	0.233	0.306
$\widehat{n}_{JT}$	1	4	5	7	19
Final height					
$\widehat{P}$	0.400	0.300	0.257	0.200	0.306
$\widehat{n}_{JT}$	2	3	4	6	19

Panse (1940) showed how the variance of the genetic variance among F_3 families can be used to estimate n_e. At the *i*th locus, the expected value of this quantity is $(1/2)(\alpha_i^2/2)^2 - (\alpha_i^2/4)^2 = \alpha_i^4/16$. Summing over all loci, and assuming equal effects, this value becomes $\sigma^2(\sigma_A^2) = n\alpha^4/16$. Thus, the ratio of the squared mean genetic variance within F_3 families to the variance of those variances,

$$\widehat{n}_P = \frac{(\overline{\sigma_A^2})^2}{\sigma^2(\sigma_A^2)} \tag{11.38}$$

provides an estimate of the effective number of loci.

In addition to its assumption of unlinked loci with additive effects, a major difficulty with this technique is its reliance on genetic variance estimates. It is difficult to determine these with a great deal of precision and is even more difficult to procure good estimates of the variance of σ_A^2 among lines.

Literature Cited

Burton, G. W. 1951. Quantitative inheritance in pearl millet *(Pennisetum glaucum). Agron. J.* 43: 409–417. [11]

Carson, H. L., and R. Lande. 1984. Inheritance of a secondary sexual character in *Drosophila silvestris. Proc. Natl. Acad. Sci. USA* 81: 6904–6907. [11]

Castle, W. E. 1921. An improved method of estimating the number of genetic factors concerned in cases of blending inheritance. *Proc. Natl. Acad. Sci. USA* 81: 6904–6907. [11]

Cavalli, L. L. 1952. An analysis of linkage in quantitative inheritance. *In* E. C. R. Reeve and C. H. Waddington (eds.), *Quantitative inheritance,* pp. 135–144. His Majesty's Stationary Office, London. [11]

Chai, C. K. 1956. Analysis of quantitative inheritance of body size in mice. II. Gene action and segregation. *Genetics* 41: 165–178. [11]

Charlesworth, B., R. Lande, and M. Slatkin. 1982. A neo-Darwinian commentary on macroevolution. *Evolution* 36: 474–498. [11]

Choo, T. M. 1981. Doubled haploids for studying the inheritance of quantitative characters. *Genetics* 99: 525–540. [11]

Choo, T. M., and E. Reinbergs. 1982. Estimation of the number of genes in doubled haploid populations of barley *(Hordeum vulgare). Can. J. Genet. Cytol.* 24: 337–341. [11]

Chovnick, A., and A. S. Fox. 1953. The problem of estimating the number of loci determining quantitative variation in haploid organisms. *Am. Nat.* 87: 263–267. [11]

Cockerham, C. C. 1980. Random and fixed effects in plant genetics. *Theor. Appl. Genet.* 56: 119–131. [11]

Cockerham, C. C. 1986. Modifications in estimating the number of genes for a quantitative character. *Genetics* 114: 659–664. [11]

Comstock, R. E., and F. D. Enfield. 1981. Gene number estimation when mutliplicative genetic effects are assumed—growth in flour beetles and mice. *Theor. Appl. Genet.* 59: 373–379. [11]

Coyne, J. A., and R. Lande. 1985. The genetic basis of species differences in plants. *Am. Nat.* 126: 141–145. [11]

Croft, J. H., and G. Simchen. 1965. Natural variation among monokaryons of *Collybia velutipes. Am. Nat.* 99: 451–462. [11]

David, F. N. 1981. *Order statistics*, 2nd Ed. Wiley, New York, NY. [11]

Emerson, R. A., and E. M. East. 1913. The inheritance of quantitative characters in maize. *Bull. Agric. Exp. Sta. Neb.* 2. [11]

Fenster, C. B., and K. Ritland. 1994. Quantitative genetics of mating system divergence in the yellow monkeyflower species complex. *Heredity* 73: 422–435. [11]

Galen, C., J. S. Shore, and H. Deyoe. 1991. Ecotypic divergence in alpine *Polemonium viscosum:* genetic structure, quantitative variation, and local adaptation. *Evolution* 45: 1218–1228. [11]

Goodwin, R. H. 1944. The inheritance of flowering time in a short-day species, *Solidago sempervirens* L. *Genetics* 29: 503–519. [11]

Gottlieb, L. D. 1984. Genetics and morphological evolution in plants. *Am. Nat.* 123: 681–709. [11]

Gould, S. J. 1980. Is a new and general theory of evolution emerging? *Paleobiol.* 6: 119–130. [11]

Hammond, J. J., and C. O. Gardner. 1974. Modification of the variety cross diallel model for evaluating cycles of selection. *Crop Sci.* 14: 6–8. [11]

Hard, J. J., W. E. Bradshaw, and C. M. Holzapfel. 1992. Epistasis and the genetic divergence of photoperiodism between populations of the pitcher-plant mosquito, *Wyeomyia smithii. Genetics* 131: 389–396. [11]

Harrison, G. A., and J. J. T. Owen. 1964. Studies on the inheritance of human skin colour. *Ann. Hum. Genet.* 28: 27–37. [11]

Harter, H. L. 1961. Expected values of normal order statistics. *Biometrika* 48: 151–166. [11]

Harter, H. L. 1970a. *Order statistics and their use in testing and estimation. Volume 1: Tests based on range and studentized range of samples from a normal population.* U. S. Gov. Printing Off., Washington, DC. [11]

Harter, H. L. 1970b. *Order statistics and their use in testing and estimation. Volume 2: Estimates based on order statistics of samples from various populations.* U. S. Gov. Printing Off., Washington, DC. [11]

Hayman, B. I. 1960a. The separation of epistatic from additive and dominance variation in generation means. *Genetica* 31: 371–390. [11]

Hayman, B. I. 1960b. Maximum likelihood estimation of genetic components of variation. *Biometrics* 16: 369–381. [11]

Hill, J. 1966. Recurrent backcrossing in the study of quantitative inheritance. *Heredity* 21: 85–120. [11]

Hill, W. G. 1982a. Dominance and epistasis as components of heterosis. *Z. Tierzüchtg. Züchtgsbiol.* 99: 161–168. [11]

Hill, W. G., and P. J. Avery. 1978. On estimating number of genes by genotype assay. *Heredity* 40: 397–403. [11]

Jaenike, J. 1987. Genetics of oviposition-site preference in *Drosophila tripunctata. Heredity* 59: 363–369. [11]

Jinks, J. L., and J. M. Perkins. 1969. The detection of linked epistatic genes for a metrical trait. *Heredity* 24: 465–475. [11]

Jinks, J. L., and P. Towey. 1976. Estimating the number of genes in a polygenic system by genotype assay. *Heredity* 37: 69–81. [11]

Jones, C. D. 2001. Extension of the Castle-Wright effective factor estimator to sex linkage and haplodiploidy. *J. Hered.* 92: 274–276. [11]

Keightley, P. D. 1994. The distribution of mutation effects on viability in *Drosophila melanogaster. Genetics* 138: 1315–1322. [11]

Kendall, M., and A. Stuart. 1977. *The advanced theory of statistics. Vol. 1. Distribution theory*. 4th Ed. Macmillan, New York, NY. [11]

Kermicle, J. L. 1969. Androgenesis conditioned by a mutation in maize. *Science* 166: 1422–1424. [11]

Khambanonda, I. 1950. Quantitative inheritance of fruit size in red pepper *(Capsicum frutescens* L.) *Genetics* 35: 322–343. [11]

Lande, R. 1981. The minimum number of genes contributing to quantitative variation between and within populations. *Genetics* 99: 541–553. [11]

Lander, E. S., and D. Botstein. 1989. Mapping Mendelian factors underlying quantitative traits using RFLP linkage maps. *Genetics* 121: 185–199 (Correction 136: 705). [11]

Luckinbill, L. S., J. L. Graves, A. H. Reed, and S. Koetsawang. 1988. Localizing genes that defer senescence in *Drosophila melanogaster*. *Heredity* 60: 367–374. [11]

Lynch, M. 1991. The genetic interpretation of inbreeding depression and outbreeding depression. *Evolution* 45: 622–629. [11]

Macnair, M. R., and Q. J. Cumbes. 1989. The genetic architecture of interspecific variation in *Mimulus. Genetics* 122: 211–222. [11]

Mather, K., and J. L. Jinks. 1982. *Biometrical genetics.* 3rd Ed. Chapman and Hall, NY. [11]

Melchinger, A. E., and C. Flachenecker. 2006. An extension of the Smith model for quantitative genetic analysis of response under recurrent selection. *Plant Breed.* 125: 644–646. [11]

Mohamed, A. H. 1959. Inheritance of quantitative characters in *Zea mays.* I. Estimation of the number of genes controlling the time of maturity. *Genetics* 44: 713–724. [11]

Mohamed, A. H., and A. S. Hanna. 1964. Inheritance of quantitative characters in rice. I. Estimation of the number of effective factor pairs controlling plant height. *Genetics* 49: 81–93. [11]

Mulitze, D. K., and R. J. Baker. 1985a. Evaluation of biometrical methods for estimating the number of genes. 1. Effect of sample size. *Theor. Appl. Genet.* 69: 553–558. [11]

Mulitze, D. K., and R. J. Baker. 1985b. Evaluation of biometrical methods for estimating the number of genes. 2. Effect of type I and type II statistical errors. *Theor. Appl. Genet.* 69: 559–566. [11]

Nandamuri, S. P., B. E. Dalton, and K. L. Carleston. 2017. Determination of the genetic architecture underlying short wavelength sensitivity in Lake Malawi cichilds. *J. Hered.* 108: 379–390. [11]

Nitzsche, W., and G. Wenzel. 1977. *Haploids in plant breeding*. Paul Parey, Hamburg. [11]

O'Brien, S. J. (ed.) 1990. *Genetic maps*, 5th Ed. Cold Spring Harbor Press, Cold Spring Harbor, NY. [11]

Ollivier, L., and L. L. G. Janns. 1993. A note on the estimation of the effective number of additive and dominant loci contributing to quantitative variation. *Genetics* 135: 907–909. [11]

Orr, H. A., and J. A. Coyne. 1992. The genetics of adaptation: a reassessment. *Am. Nat.* 140: 725–742. [11]

Panse, V. G. 1940. A statistical study of quantitative inheritance. *Ann. Eugenics* 10: 76–105. [11]

Plomion, C., N. Bahrman, C.-E. Durel, and D. M. O'Malley. 1995. Genomic mapping in *Pinus pinaster* (maritime pine) using RAPD and protein markers. *Heredity* 74: 661–668. [11]

Pooni, H. S., J. L. Jinks, and J. F. F. de Toledo. 1985. Predicting and observing the properties of second cycle hybrids using basic generations and inbred line $\times$ F_1 crosses. *Heredity* 54: 121–129. [11]

Powers, L. 1942. The nature of the series of environmental variances and the estimation of the genetic variances and the geometric means in crosses involving species of *Lycopersicon*. *Genetics* 27: 561–575. [11]

Powers, L. 1951. Gene analysis by the partitioning method when interactions of genes are involved. *Bot. Gaz.* 113: 1–23. [11]

Ryder, E. J. 1958. The effects of complementary epistasis on the inheritance of a quantitative character, seed size in lima beans. *Agron. J.* 50: 298–301. [11]

Sarhan, A. E. and B. G. Greenberg. 1962. *Contributions to order statistics*. Wiley, New York, NY. [11]

Schnell, F. W., and C. C. Cockerham. 1992. Multiplicative vs. arbitrary gene action in heterosis. *Genetics* 131: 461–469. [11]

Smith, H. H. 1937. The relation between genes affecting size and color in certain species of *Nicotiana*. *Genetics* 22: 361–375. [11]

Smith, O. S. 1979a. A model for evaluating progress from recurrent selection. *Crop Sci.* 19: 223–226. [11]

Smith, O. S. 1979b. Application of a modified diallel analysis to evaluate recurrent selection for grain yield in maize. *Crop Sci.* 19: 819–822. [11]

Smith, O. S. 1983. Evaluation of recurrent selection in BSSS, BSCB1, and BS13 maize populations. *Crop Sci.* 23: 35–40. [11]

Snape, J. W., A. J. Wright, and E. Simpson. 1984. Methods for estimating gene numbers for quantitative characters using doubled haploid lines. *Theor. Appl. Genet.* 67: 143–148. [11]

Sprague, G. F., and B. Brimhall. 1949. Quantitative inheritance of oil in the corn kernel. *Agron. J.* 41: 30–33. [11]

Templeton, A. R. 1977. Analysis of head shape differences between two interfertile species of Hawaiian *Drosophila*. *Evolution* 31: 330–341. [11]

Towey, P., and J. L. Jinks. 1977. Alternative ways of estimating the number of genes in a polygenic system by genotype assay. *Heredity* 39: 399–410. [11]

Vallejos, C. E., N. S. Sakiyama, and C. D. Chase. 1992. A molecular marker-based linkage map of *Phaseolus vulgaris* L. *Genetics* 131: 733–740. [11]

Waters, N. F. 1931. Inheritance of body weight in domestic fowls. *Rhode Island Agric. Exp. Sta. Bull.* 228: 7–103. [11]

Wehrhahn, C. and R. W. Allard. 1965. The detection and measurement of the effects of individual genes involved in the inheritance of a quantitative character. *Genetics* 51: 109–119. [11]

Wilkens, H. 1971. Genetic interpretation of regressive evolution processes: studies on hybrid eyes of two *Astyanax* cave populations (Characidae, Pisces). *Evolution* 25: 530–544. [11]

Wright, S. 1968. *Evolution and the genetics of populations. I. Genetic and biometric foundations.* Univ. Chicago Press, Chicago. [11]

Wu, C.-I., and M. F. Palopoli. 1994. Genetics of postmating reproductive isolation in animals. *Ann. Rev. Ecol. Syst.* 27: 283–308. [11]

Zeng, Z.-B. 1992. Correcting the bias of Wright's estimates of the number of genes affecting a quantitative character: a further improved method. *Genetics* 131: 987–1001. [11]

Zeng, Z.-B., D. Houle, and C. C. Cockerham. 1990. How informative is Wright's estimator of the number of genes affecting a quantitative character? *Genetics* 126: 235–247. [11]